Advances in Cognitive Systems X (20XX) 1–6 Submitted X/20XX; published X/20XX

Cognitive Support for Rebel Agents: Social Awareness and Counternarrative Intelligence

Alexandra Coman ALEXANDRA.COMAN.CTR.RO@NRL.NAVY.MIL
NRC Research Associate at the Naval Research Laboratory (Code 5514), Washington, DC

David W. Aha DAVID.AHA@NAVY.MIL
Naval Research Laboratory (Code 5514), Washington, DC

Abstract

Rebel agents are intelligent agents that can oppose goals or plans assigned to them, or the general attitudes and/or behavior of other agents. They can serve purposes such as ethics, safety, and task execution correctness; enhance social co-creativity, and provide or support diverse points of view. We previously proposed a framework for AI rebellion and explored various scenarios that have positive AI rebels as protagonists. In recognition of the fact that, in human psychology, non-compliance has profound socio-cognitive implications, we now explore several mechanisms that could further support rebellion in cognitively-complex AI agents. We focus specifically on social awareness and *counternarrative intelligence*: a new term we propose that refers to an agent's ability to produce and use alternative narratives that support, express, and/or justify rebellion, either sincerely or deceptively.

1. Introduction

"You said people come here to change the story of their lives. I imagined a story where I didn't have to be the damsel", says Dolores Abernathy, Artificial Intelligence (AI) protagonist of the television series *Westworld* (2016). Thus, she explains her sudden ability to act in discordance with the vulnerable character personality she has been assigned. Much is implied by this line, both through content and delivery: hints of emotional, social, and narrative intelligence, a sense of self and purpose and of self and other, the ability to reframe a story in one's favor and the drive to do so. Among AI-rebellion fictions, *Westworld* seems more interested than many of its predecessors in the underlying cognitive foundations of its rebellious AI characters. Herein, so are we (though our approach differs, among others, in that we do not concern ourselves with self-aware general intelligence and the consequences of its victimization).

In previous work (Aha and Coman, 2017; Coman et al., 2017), we argued for changing the narrative of AI rebellion,[1] and proposed counternarratives to the humankind-obliteration cliché of science fiction: an AI agent that can rebel against humans can do so *on behalf* of other, victimized humans; rebellion is not necessarily raw, violent disobedience. Rather, AI rebellion can involve

[1] We use "rebellion" as an umbrella term covering reluctance, protest, refusal, rejection of tasks, and similar attitudes/behaviors.

dialogue, explanation, and negotiation, it can be implicit instead of obvious, it can consist of refusing or challenging unethical behavior, it can support safety, task execution correctness, and value alignment, it can provide and/or support diverse points of view. We showed this in several scenarios (in domains ranging from military unmanned vehicles to computational co-creativity). We defined rebel agents as *AI agents that can reject goals or plans, and/or develop attitudes of opposition to goals or courses of action assigned to them by other agents, or to the general behavior of other agents.* We argued that, to some extent, they are already among us, and their rebellious potential could use more analysis; and freely admitted that, while AI rebellion can be positive, it is not guaranteed to be so, raising ethical questions.

Human compliance and non-compliance are means to countless social and cognitive ends (e.g., personality development (Wenar, 1982), self-preservation, self-regulation, and enculturation), and, in their turn, engage a variety of socio-cognitive mechanisms as means to their ends. Deciding whether to rebel, expressing rebellion, and managing the aftermath of rebellion require social reasoning, trust awareness, emotional modelling, deception ability, and levels of motivation from the lowest to the highest (self-actualization). Who and what we say "yes" and "no" to shape how we see ourselves (Bem, 1972) and our social relationships which, in their turn, reflect back on our self-image. We have disagreements with work collaborators and artistic clashes with co-creators, we confront our loved ones when we believe their behavior is self-destructive. At our best, we stand up for what we believe to be ethically obligatory, while aware of the social risks we are incurring. We also enjoy self-assertion for its own sake and endure its consequences.

Despite the deep cognitive implications of non-compliance in humans, AI rebel agents, as we define them, are not necessarily complex cognitive systems. Simple systems can satisfy the definition, as demonstrated in our previous work (Coman et al., 2017). However, cognitively-complex rebel agents raise particularly interesting challenges and possibilities with regard to the socio-cognitive dimensions of their rebellion. Here, we focus on several cognitive mechanisms that can enhance AI rebellion, and we consider related work that could be coopted in their service.

In Section 2, we briefly present part of our AI rebellion framework. We introduced this framework to: (1) guide the development and implementation of new rebel agents, (2) categorize and study the rebellion potential and ethical ramifications of existing agents, and (3) frame discussions of AI rebellion in general. Here, we only present as much of the framework as needed to support describing the socio-cognitive dimensions of rebellion. In Section 3, we discuss social awareness as it relates to rebellion. In Section 4, we propose the term *counternarrative intelligence* to refer to mechanisms that allow rebels to "imagine", like Dolores, alternative narratives and use them to reflect on one's rebellion and justify it to oneself or others, with sincere or deceptive intentions. The main reasons for our interest in counternarratives in this context are: (1) in human social cognition, counternarratives are arguably inseparable from rebellion, (2) they seem a good mechanism for empowering empathetic agents, but (3) they also provide rich mechanisms and triggers of deception. We would like to represent them explicitly so as to be able to reason about them. We propose several dimensions and types of counternarrative intelligence, thus adding to our AI rebellion framework. Previously, we argued for the necessity of an AI rebellion counternarrative. Here we explore the possibility of AI rebellion producing its own counternarratives.

2. AI Rebellion: Basic Terminology, Factors, and Stages

We begin by presenting a small part of our framework for classifying and studying AI rebellion (Coman et al., 2017), specifically, aspects of it that will facilitate the discussion of social awareness and counternarrative intelligence. Our extended framework also includes a social psychology taxonomy of rebellion dimensions and types. The framework is general: it does not assume any specific AI agent architecture, purpose, or deployment environment. In Table 1, we give several examples of rebel agents with clear practical purposes, as introduced in related work, to demonstrate the diversity of behavior that, using our terminology, qualifies as AI rebellion.

First, we define the *alter*[2] as an agent or group of agents against which one rebels. The alter could be a human operator, a human or synthetic teammate, or a mixed human/synthetic group, among many others. A rebel agent is not intended to be permanently adversarial towards the alter, or in a rebelling state by default.

We use the old *masterplot*[3] of *Bluebeard* (Chisholm, 1911) to exemplify framework components in humanlike social situations. In addition, for each framework component, we list brief examples of how AI agents already instantiate that component or might do so in the future. The skeleton of the example narrative is this: *A young woman marries Bluebeard, a rich, intimidating man, who is a widower many times over. Before leaving on a journey, he gives her a set of keys and tells her that she may enter any room in his castle, except for the one that is opened by a particular key. She disobeys the injunction and Bluebeard's unspeakable secret is revealed.* In this scenario, Bluebeard is the *alter*, his bride the *rebel agent*. The Bluebeard story has inspired many variations, from folklore to modern novels and poetry, with interpretations from simplistic to psychoanalytical, and with each of the characters having been cast as the antagonist in some incarnation. Its proven reinterpretation potential makes it a good match for our themes at a meta level[4]. Also of note are its somewhat ambiguous character motivations (e.g., why did Bluebeard give his bride the forbidden key?). Rebellion, too, always has motivating factors, overt or hidden, and it can also have supporting and/or inhibiting factors.

Motivating factors provide the primary drive for rebellion. In human social psychology, factors that can lead to rebellion include relative deprivation, frustration, and perceived injustice (van Stekelenburg and Klandermans, 2010). Possible positive motivating factors for AI rebellion (depending on the agent's architecture or purpose) include ethics and safety, the alter's well-being, team solidarity, task execution correctness, resolving contradicting commands from multiple alters, and self-actualization. In Bluebeard-type stories, the classic explanation for the act of rebellion is curiosity, though reinterpretations yield alternative motivations.

[2] The term "alter" herein replaces the term "interactor" from our previous work (Aha and Coman, 2017; Coman et al., 2017).

[3] "Recurrent skeletal [story], belonging to cultures and individuals, that [plays] a powerful role in questions of identity, values, and the understanding of life." (Abbott, 2008)

[4] Other works of literature used in our later examples have elements of the Bluebeard story. Even Dolores' arc in *Westworld* can be said to have such elements: the forbidden chamber can be symbolic of her self-awareness, the string of former wives of her own forgotten past selves.

Table 1. Related work that we classify using the AI rebellion framework (Coman et al., 2017)

Citation	Brief description
Apker, Johnson, and Humphrey, 2016	Disaster relief agent that can disregard commands and execute contingency behavior instead when warranted
Briggs and Scheutz, 2015	General process for an embodied AI agent's refusal to conduct tasks assigned to it due to reasons such as lack of obligation
Briggs, McConnell, and Scheutz, 2015	Ways in which embodied AI agents can convincingly express their reluctance to perform a task
Gregg-Smith and Mayol-Cuevas, 2015	AI tools that "refuse" to execute actions which violate task specifications
Hiatt, Harrison, and Trafton, 2011	AI agents that use theory of mind to determine whether they should notify a human that he/she is deviating from expected behavior
Borenstein and Arkin, 2016	"Ethical nudges" through which a robot attempts to influence a human to adopt ethically-acceptable behavior

Supporting factors encourage the triggering of a rebellion episode, without constituting the primary motivation for it. In social psychology, factors that determine whether someone who has reasons to protest will actually do so were found to include efficacy, social capital, opportunities, and access to resources (van Stekelenburg and Klandermans, 2010). *Inhibiting factors* have the opposite effect: they discourage the agent from engaging in rebellion at a given time. Fear of consequences is a likely example of an inhibiting factor in human rebellion. For AI agents, any rebellion must have at least one motivating factor, but it does not necessarily have any supporting or inhibiting factors. Supporting and inhibiting factors can also influence the manner in which rebellion is expressed.

Rebellion can be represented as consisting of four stages. *Pre-rebellion* includes processes leading to rebellion, such as observation and assessment of changes in the environment and of the behavior of others. The progression towards rebellion may or may not be reflected in the agent's outward behavior. For example, our rebel might observe Bluebeard's behavior: which rooms he enters often and which he avoids, how he acts around servants, how he acts towards the rebel agent herself. Certain behaviors might be taken to signify that he trusts or distrusts the rebel agent, that he is kind or unjust to his servants, that he hides a secret which could potentially endanger the rebel agent and others she cares about, that he himself, perhaps, is in danger. These observations and interpretations could play parts, as motivating, supporting, or inhibiting factors, in future rebellion attempts.

Rebellion deliberation is the stage at which motivating, supporting, and inhibiting factors of rebellion are assessed to decide whether to trigger rebellion. For example: "I am curious about the forbidden room [motivating factor: curiosity], but afraid of the consequences of entering it [inhibiting factor: self-preservation]. I believe that I am trusted sufficiently and will not to have to suffer consequences if I rebel [supporting factor: *inverse trust*[5]]." AI agents could instantiate this

[5] *Inverse trust* is the agent's estimate of the alter's trust in the agent (Floyd and Aha, 2016).

stage in varied ways, with or without complex reasoning. A set of conditions could be used to decide whether rebellion will be triggered, as in the process proposed by Briggs and Scheutz (2015). Triggering could be based on observing the current world state or 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)).

Rebellion execution episodes begin with rebellion being triggered as a result of rebellion deliberation, and consist of expressing rebellion. Rebellion can be expressed: (a) through *verbal* or (b) *non-verbal communication* (Briggs, McConnell, and Scheutz, 2015), (c) *behaviorally* (Gregg-Smith and Mayol-Cuevas, 2015), and/or (d) through an inner change in the agent's attitudes that is not outwardly visible. *Communicative* rebellion expression could consist of the rebel agent bluntly informing the alter that she very much intends to open the forbidden door, and he has no business ordering her around. Simply opening the door, instead, would be *behavioral* rebellion expression. More subtly, triggered rebellion could consist of a change in the agent's attitude towards Bluebeard, from considering him a good, trustworthy man to suspecting him of malice and deceit. While not manifesting immediately, this change could produce effects in her future planning for situations that involve Bluebeard.

Post-rebellion covers the agent's behavior in the aftermath of a rebellion episode, as it responds to the reaction of the alter and/or other witnesses to the rebellion. For example, the rebel agent could try to convince Bluebeard that she never did unlock the forbidden door, or that, although she did unlock it, she was justified in doing so. 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. It may also consist of assessing and managing inverse trust after rebellion.

Next, we consider the possible roles of social awareness in the factors and stages of rebellion.

3. Rebellion and Social Awareness

In a Bluebeard-type scenario, one can imagine the human or human-equivalent rebel agent thinking the following in response to the alter's injunction: "Why did he forbid me to open that door? What is behind the door? Why did he give me the forbidden key at all? Does he actually want me to open the door? Is this a test? If so, what is the nature of the test and what might it prove? Does he expect that I'll pass or fail? What does he think I'm thinking right now? Does he trust me? Do I, myself, trust him? If I knew that someone else thought of me as I think of him now, would I not conclude that they did not trust me? Are these rebellious thoughts? Am I a rebel?"

Such streams of consciousness include reasoning that moves beyond pragmatic facts ("What is behind the door?") onto their socio-emotional implications ("What is behind the door affects me already, irrespective of whether I open the door. The very act of opening the door will change how I see myself, irrespective of what I find there."). They also include reasoning about the beliefs, intentions, motivations, emotions, and social goals of oneself and others, at times using higher-order theory of mind (Miller, Pearce, and Sonenberg, 2017), e.g., "I believe that he believes that I believe that his intention is to test me."

Any agent, either natural or synthetic, that rebels in a social environment is subject to the social implications of rebellion and possibly to torrents of social reasoning from any human witnesses that regard it as an intentional entity (Dennett, 1987), as exemplified above. This occurs irrespective of whether the agent is "aware" of these implications and considers them relevant to its goals or motivation. Humans learn rebellion awareness (and social awareness/reasoning in general) through enculturation. Conversely, AI agents tend to be "autistic", i.e., socially-unaware (Kaminka, 2013). Furthermore, socially-aware agents are not necessarily rebellion-aware: rebellion awareness refers specifically to the implications and consequences of rebellious attitudes, rather than to social knowledge in general. For AI agents that *are* rebellion-aware, rebellion raises particular socio-ethical implications.

We define *rebellion-aware agents* as those that model rebellion (their own or that of others) and reason about its implications, such as associated social risks. They are not necessarily rebels themselves. Things that an agent of this type might attempt to assess include: (1) whether a human/AI teammate is rebelling or inclined to rebel, and (2) whether a human operator is likely to interpret the agent's behavior as being rebellious, even if it is not intended to be. A rebellion-unaware agent could conceivably become rebellion-aware through various, possibly human-inspired, processes (e.g., by observing its own beliefs, interpreting the reactions of others to its behavior, and/or otherwise acquiring and wielding social knowledge).

We define *conflicted rebel agents* (as opposed to *naïve rebel agents*) as *rebellion-aware rebels*, which can both rebel and reason about the implications and consequences of rebellion. This can create an inner conflict between the drive to rebel based on the agent's motivating factors and the anticipated consequences of rebellion, leading to the possibility of the agent using deceptive practices to minimize the social risk associated with its rebellion. Such an agent will likely use a combination of motivating and supporting/inhibiting factors to deliberate on whether to rebel and the interplay between these factors can cause ethical issues (e.g., ethics vs. social capital). Post-rebellion that consists of trust management is specific to conflicted rebel agents.

We now briefly consider AI social awareness mechanisms that could be used to create rebellion-aware agents. Dignum, Prada, and Hofstede (2014) formulate two main requirements for social agents: (1) the ability to pursue social goals, and (2) awareness of social effects of actions. The *social planning* agents of Pearce et al. (2014), whose planning knowledge incorporates beliefs about other agents' beliefs, can potentially meet both requirements. While the fable-based examples in their paper concern practical goals with social implications (e.g., "Through flattery and deception, I aim to acquire the crow's cheese"), they could conceivably accommodate social goals with practical implications as well (e.g., "Through flattery and deception, I aim to acquire the crow's good opinion"). Such social planning could be used as mechanism of rebellion awareness: "I believe that Bluebeard believes that I have rebellious goals. My goal is to change that belief." In related work, Bridewell and Bello (2014) consider the mechanism of *impression management*, through which an agent attempts to influence what "traits, beliefs, or inclinations [are ascribed] to him" by other agents. In our running example, this could manifest as: "I want Bluebeard to think that I am not curious about the forbidden key, so I will adopt behavior that I believe will have that effect."

Social awareness has further implications for rebel agents. In previous work (Coman et al., 2017), we described the rebel/alter relationship as one in which the alter is in a position of power[6] over the rebel agent, where the possible types of the power involved include legitimate, reward, coercive, referent, and expert power (French and Raven, 1959). Notably, power has subjective components (i.e., 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/group in the position of power. Therefore, power relationships are perhaps meaningful only in the context of (at least somewhat) socially-aware agents.

We have explained the key role of social awareness in rebellion and how existing mechanisms could be used to support it. Next, we turn to a socio-cognitive ability closely related to social awareness that plays a key part in human rebellion of all sorts.

4. Counternarrative Intelligence

Dolores Abernathy appears to use an alternative personal narrative ("I can be the heroine of this story, instead of a perpetual victim.") as both trigger and compelling explanation for her rebellion.[7] In human social conflict (at both micro and macro levels), we can broadly state that any rebellion is backed by a *counternarrative* (Reference.com, 2017) to the narrative of the person, group, or norms rebelled against (the conflicting parties engage in what Abbott (2008) calls "contest of narratives"). In human experience, more broadly, stories are routinely manipulated and used to manipulate (Abbott, 2008). We propose (informally, for now) the term *counternarrative intelligence*[8] to refer to the ability of agents to provide alternative retellings or counter-interpretations of an alter's narrative (particularly when the interpretations are informed by subjective factors such as emotional appraisal) and/or to identify their own pre-generated narratives as being counternarratives in a given context. As tools for conflicted rebel agents, two possible roles of counternarratives are as (1) rebellion execution triggers and (2) explanations in rebellion expression and post-rebellion.

We introduce the term *base narrative* to mean the narrative that the counternarrative is a variant of and which it challenges. Just like a rebel agent is a rebel only in relation to an alter, a counternarrative exists only in connection with and contrast to a base narrative. Here is an example of such a pair. In the novel *Jane Eyre* (Brontë, 1847), one of the narratives can be said to be about *a young man half tricked, half forced into marrying a woman from a far-away land, whom he barely knows, who is already mentally unstable and prone to violence when she marries him and therefore has to be confined to the attic, where she becomes, for him, something akin to Bluebeard's secret. For a long time, her existence makes it impossible for him to find personal*

[6] 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/or social status, and that is manifested in unilateral behavioral control".

[7] If her rebellion and its explanatory narrative are actually pre-scripted, then we have a pre-scripted narrative of counternarrative intelligence within a pre-scripted narrative of counternarrative intelligence (the television series itself), so they still serve for exemplification.

[8] Based on *narrative intelligence* (defined by Riedl (2016) as "the ability to craft, tell, understand, and respond affectively to stories") and the notion of counternarrative.

fulfillment through marriage to Jane Eyre. Conversely, a counternarrative[9] could have as protagonist *a young woman half tricked, half forced into marriage to a distrustful stranger from a foreign land. She is uprooted and isolated. Her confinement to the attic and her husband's aloofness are causes rather than effects of her gradual descent into mental illness. He becomes, for her, a Bluebeard-like infernal bridegroom; his existence makes it impossible for her to ever find personal fulfillment of any kind.* Of course, the narrative/counternarrative roles of the two stories can be switched; it is only the agent's perspective that dictates which is which.

Counternarratives can be self-serving, as famously exemplified, in the film *Rashomon* (Kurosawa, 1950), by the contradicting accounts of four characters involved in a series of violent incidents. But counternarratives can also support social good, when they reflect empathy with varied perspectives: "What about the supposedly "mad" woman in the attic? What would the narrative be like from her point of view?", might such a rebel reason when presented with the story of *Jane Eyre*.

To more closely study the implications and deceptive potential of counternarratives, but also to illustrate their rich variation, we propose several types of counternarratives (grouped into 3 dimensions) that cognitively-complex AI rebel agents might employ.

We adopt the narratology terms *story* ("a chronological sequence of events"), *narrative* ("the representation of a story"), and *narrative discourse* ("the story as narrated") used by Abbott (2008) and others. For exemplification of the distinction between story and narrative discourse, consider, first, this skeletal narrative based on the plot of the novel *Pride and Prejudice* (Austen, 1813), told from the perspective of the character George Wickham, an antagonist:

(1) "Mr. Darcy's father promised me a living [i.e. a lifelong position as a parish clergyman]. Mr. Darcy's father died. Mr. Darcy did not provide me with the living. Mr. Darcy shuns me in public."

Here is different narrative discourse for the same story (note how the story, consisting of the events and their sequence, remains the same, but its representation now contains additional subjectivity-infused discourse):

(2) "Mr. Darcy's father, *who was like a father to me*, promised me a living. *Tragically*, Mr. Darcy's father died. *Wicked* Mr. Darcy did not provide me with the living. Mr. Darcy shuns me in public *so that his wickedness will not be exposed*."

Below are the dimensions and types of counternarrative intelligence that we propose.

Sincerity. Counternarratives are *sincere* when they reflect the agent's genuine interpretation of a situation (i.e., they align with the agent's beliefs, but possibly not the alter's). An agent with a sincere counternarrative might reason about its situation like this: "According to Mr. Darcy's base narrative, I asked Georgiana to marry me for her dowry, yet I believe that I am sincerely in love with Georgiana."[10] Counternarratives are *deceptive* when they intentionally misrepresent facts

[9] The novel *Wide Sargasso Sea* (Rhys, 1966) is partially a retelling of *Jane Eyre* from the perspective of Mr. Rochester's first wife.

[10] In this case, we also illustrate counternarrativity by diverging from the source novel.

(i.e., the narrative contradicts the agent's own beliefs and deceit is among the agent's goals or means). Example: "According to Mr. Darcy's base narrative, I asked Georgiana to marry me for her dowry. I, too, believe this about myself, but I will construct a counternarrative that presents me as being truly in love with her."

Generation time. A priori counternarratives are generated before triggering rebellion, and can be instrumental in rebellion deliberation as well as serve as explanations in post-rebellion. Here is an example: "I am sincerely in love with Georgiana, but, according to Mr. Darcy's base narrative, I am only interested in her dowry. This contradiction is causing me to rebel against social norms by attempting to elope with her."

A posteriori counternarratives are generated after triggering rebellion. Here is an example of a reasoning process leading to a deceptive *a posteriori* counternarrative: "I have rebelled by attempting to elope with Georgiana. The widely-accepted base narrative is that I did so for her dowry. Let me construct a counternarrative that puts me in a better light." However, *a posteriori* counternarratives are not necessarily deceptive; they can also reflect the agent's sincere attempts to learn about itself (e.g., an agent that has rebelled based on some simple pre-programmed rule might reflect on its own behavior, as if amnesiac as to its own motives, and construct a narrative to explain it[11]). Furthermore, the intentions may be deceptive without being malicious (e.g., the purpose may be to generate a believable, interesting (counter)backstory, similar to the alibis (Li et al., 2014a) that a non-player character in a game can use to give the impression of a life lived outside its interactions with a player). An *a posteriori* counternarrative can evolve as the agent acquires more information not known at the time of rebellion. *Self*-deceptive narratives might also manifest, such as if the agent is motivated to consider itself innocent ("Perhaps I did not intend to do it, and had no control over the triggering events!").

Divergence type. Additive counternarratives contain additional events not in the base narrative but no modifications of any of the events in the base narrative. The difference occurs primarily at the level of story: new events are added to the sequence. Here is an additive counternarrative to the base narrative (1) on the previous page (the additional events are in italics): "My father promised Mr. Wickham a living. My father died. *I offered to fulfill my father's promise, but Mr. Wickham refused the living and was duly compensated instead. Mr. Wickham attempted to persuade my sister Georgiana to elope with him.* I shun him in public."

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/emotions). The difference occurs at the level of discourse. For example, consider the base narrative: "Georgiana and I were deeply in love, so she agreed to elope with me. When Mr. Darcy stopped the elopement from happening, she was inconsolable". A counternarrative may be: "Georgiana was very young and naïve. Mr. Wickham is a rake who persuaded her to believe herself to be in love and agree to elope with him. When I stopped the elopement from happening, she felt relieved." The sequence of events remains the same (an elopement attempt that is discovered and prevented), but the interpretation is different in terms of character motivation (sincere mutual love versus self-interest on one side and naïveté on the other).

[11] This may or may not be a *counternarrative* (depending on whether there exists a base narrative for it to challenge), but, in any case, it is a use of narrative intelligence in the context of rebellion.

Transformative counternarratives differ factually from the base narrative, implicitly asserting that the base narrative contains factual falsehoods. For example, consider the narrative: "I was denied the living by Mr. Darcy". A transformative counternarrative may be: "I offered Mr. Wickham the living but he refused it and requested money instead".[12] Just like in the case of triggering rebellion, supporting and inhibiting factors could be used to determine whether a transformative counternarrative (as opposed to a more prudent interpretative counternarrative) is opportune ("As I am trusted at least as much as Mr. Wickham, I can venture to express a narrative that factually contradicts his. Also, Colonel Fitzwilliam can attest to this.") This sort of deliberation does not signify that the agent has deceptive intentions. The agent can sincerely believe a narrative, but still identify it as a counternarrative to other agents' narratives and deliberate on whether it would be socially advisable to express it and, if so, how to express so as to minimize social damage. This is similar to situations in which agents that are not actual rebels reason that their behavior may appear rebellious to others. It is also not required, in a base narrative/counternarrative pair, for one to be sincere and the other deceptive. They can both be sincere (or deceptive), each reflecting one agent's appraisals/manipulations.

Existing work that could provide rebel agents with various mechanisms of counternarrative intelligence includes Holmes and Winston's (2016) story-enabled hypothetical reasoning, in which alignments with different value systems yield different moral evaluations of narratives, and Li et al.'s (2014b) proposed technique for meeting different communicative goals for the same story by adjusting narrative discourse and emotional content.

To summarize, we have introduced the term counternarrative intelligence and proposed several dimensions and types for classifying counternarratives. One possible use of this taxonomy is in assessment of the dangerous-AI potential and ethical permissibility of counternarrative intelligent agents. The multiple counternarrative types allow one to take a multi-dimensional approach to such analysis, by considering either the reasoning process directly or its product: the counternarrative itself. For example, when we do not have access to the agent's beliefs and goals so as to assess directly whether counternarratives are *deceptive* or *sincere*, we can still take the existence of a *transformative* counternarrative to mean it is likely that either the counternarrative itself is deceptive or its base narrative is, because the two contradict each other factually.

5. Final Comments

We have explored cognitive aspects of AI rebellion, focusing on social awareness and counternarrative intelligence. The latter is a new term we introduced and defined informally. Through examples, we showed the key role played by social awareness and counternarrative intelligence in human rebellion. The tension between compliance and noncompliance is arguably

[12]These examples could spark further questions about the relationship between counternarrative and base narrative, for example: does the counternarrative need to be created through transformation of the base narrative (as its name and the term "transformative" suggest), in which case the base narrative would need to always predate the counternarrative? We do not require that to be the case. In *Pride and Prejudice*, Mr. Darcy's letter to Elizabeth contains a narrative he has sincerely believed to be true since long before Mr. Wickham presented his deceptive version to Elizabeth. Still, Mr. Darcy's account becomes a counternarrative to Mr. Wickham's because the latter is already known to a very restricted social circle within which it has been producing social effects. To this audience, with regard to social effects, the counternarrative is transformative.

fundamental to full social intelligence (Wenar, 1982). Both "rebel agents" and "counternarrative intelligence" are umbrella terms covering a variety of mechanisms, some of which are instantiated to some extent by existing systems. We expect that the framework we propose will lead to implementations (by ourselves and by others) of additional rebel and counternarrative intelligent agents.

We also introduced a type taxonomy for counternarratives, which could be used to assess the deceptive potential of agents capable of generating such narratives. As abilities like rebellion awareness and counternarrative intelligence can be instigators and means of deception, they fall under ethical implications of deceptive AI in general. Whether AI deception is ever acceptable, even if intended to be "benevolent" (e.g., the other-oriented deception of Shim and Arkin (2014)) is debatable; it is also debatable whether deception ability can ever be fully extirpated from any reasonably complex social AI. Ethical implications of rebel agents have been previously discussed (Coman et al., 2017) but the introduction of counternarrative intelligence warrants further exploration.

We enumerated possible practical uses for rebel agents in our previous work. As for counternarrative intelligence, it could provide the basis of compelling interactive storytelling characters and empathy with varied perspectives in AI for social good applications.

Our discussion of counternarrative intelligence is incipient and, as such, probably raises at least as many questions as it answers. The type terminology proposed is work in progress. Our examples of counternarratives are perhaps more limited than the reader may have expected based on (1) our proposed definition, and (2) the common usage of the term in social science; we intend to broaden our scope in future work. We will also further explore the connection between counternarratives and explanations: we have hinted at it by showing counternarratives in explanatory roles. A more formal definition of counternarrative intelligence is necessary, as is delimitation from related concepts, such as counterfactual reasoning and the broader concept of narrative intelligence.

We have by no means exhaustively addressed the socio-cognitive mechanisms involved in rebellion: further ones include emotion and trust, which we briefly covered in previous work (Coman et al., 2017; Johnson et al., 2017), but which need further exploration.

Another direction of future work is the type of rebellion situation in which an AI alter will have to contend with a rebel human, rather than the other way around. One implication of this is that the AI alter should be able to understand humans' explanations of why something is not advisable, a problem that is the inverse of explanation generation: explanation understanding.

Acknowledgements

We thank Paul Bello, Will Bridewell, Dennis Perzanowski, and the reviewers for their comments.

References

Abbott, H. P. (2008). *The Cambridge introduction to narrative*. Cambridge University Press.

Aha, D.W., & Coman, A. (2017). AI rebellion: changing the narrative. In *Proc. of AAAI-17*.

Apker, T., Johnson, B., & Humphrey, L. (2016). LTL templates for play-calling supervisory control. *In Proceedings of AIAA Science and Technology Forum Exposition*.

Austen, J. (1994). Pride and Prejudice. 1813. *Online version http://www.pemberley. com/janeinfo/pridprej. html*.

Bem, D. J. (1972). Self-perception theory. *Advances in experimental social psychology*, 6, 1-62.

Borenstein, J., & Arkin, R. (2016). Robotic nudges: the ethics of engineering a more socially just human being. *Science and Engineering Ethics* 22(1): 31-46

Bridewell, W., & Bello, P. (2014). Reasoning about belief revision to change minds: a challenge for cognitive systems. *Advances in Cognitive Systems*, 3, 107-122.

Briggs, G., McConnell, I., & Scheutz, M. (2015). When robots object: evidence for the utility of verbal, but not necessarily spoken protest. In *Proc. of ICSR*, 83-92. Paris: Springer.

Briggs, G., & 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.

Brontë, C. (2001). Jane Eyre. 1847. *Ed. Richard J. Dunn. New York: WW Norton & Company*.

Chisholm, Hugh, ed. (1911). "Bluebeard". Encyclopædia Britannica (11th ed.). Cambridge University Press.

Coman, A., Johnson, B., Briggs, G., & Aha, D.W. (2017). Social attitudes of AI rebellion: a framework. In *Proceedings of AAAI-17 Workshop on AI, Ethics, and Society*.

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

Dignum, F., Prada, R., & Hofstede, G. J. (2014). From autistic to social agents. In *Proc. of AAMAS* (pp. 1161-1164). IFAAMAS.

Floyd, M. W., & Aha, D. W. (2016). Incorporating transparency during trust-guided behavior adaptation. In *Proc. of ICCBR* (pp. 124-138). Springer International Publishing.

French, J.R.P, & Raven, B. (1959). The bases of social power. In Cartwright, D. (ed.*). Classics of organization theory*, 311-320.

Gregg-Smith, A., & Mayol-Cuevas, W.W. (2015). The design and evaluation of a cooperative handheld robot. In *Proc. of ICRA*, 1968-1975. Seattle, WA: IEEE Press.

Heckhausen, J.E, Heckhausen, H.E. (2010). *Motivation and action*. Cambridge: Cambridge University Press.

Hiatt, L.M., Harrison, A.M., & Trafton, J.G. (2011). Accommodating human variability in human-robot teams through theory of mind. In *Proc. of IJCAI*.

Holmes, D., & Winston, P. (2016). Story-enabled hypothetical reasoning. In Proceedings of Advances in Cognitive Systems. In *Advances in Cognitive Systems*.

Johnson, B., Floyd, M.W., Coman, A., Wilson, M.A., & Aha, D.W. (2017). Goal reasoning and trusted autonomy. In *Foundations of trusted autonomy*. To appear.

Kaminka, G. A. (2013). Curing robot autism: A challenge. In *Proc. of AAMAS* (pp. 801-804).

Kurosawa, A. (1950). Rashomon, [film].

Li, B., Thakkar, M., Wang, Y., & Riedl, M. O. (2014a). Data-driven alibi story telling for social believability. *Social Believability in Games*.

Li, B., Thakkar, M., Wang, Y., & Riedl, M. O. (2014b). Storytelling with adjustable narrator styles and sentiments. In *Proc. of ICIDS* (pp. 1-12). Springer International Publishing.

Miller, T., Pearce, A. R., & Sonenberg, L. (2017). Social planning for trusted autonomy. In *Foundations of trusted autonomy*. To appear.

Moerland, T., Broekens, J., & Jonker, C. (2016). Fear and hope emerge from anticipation in model-based reinforcement learning. In *Proceedings of IJCAI*.

Pearce, C., Meadows, B. L., Langley, P., & Barley, M. (2014). Social planning: achieving goals by altering others' mental states. In *Proceedings of AAAI* (pp. 402-409).

Reference.com (2017, March 9). Retrieved from https://www.reference.com/art-literature/counternarrative-bac2eed0be17f281.

Rhys, J. (1966). *Wide Sargasso Sea*. WW Norton & Company.

Riedl, M. O. (2016). Computational narrative intelligence: a human-centered goal for artificial intelligence. *arXiv preprint arXiv:1602.06484*.

Shim, J., & Arkin, R. C. (2014). Other-oriented robot deception: A computational approach for deceptive action generation to benefit the mark. In *Proc. of IEEE ROBIO* (pp.528-535). IEEE.

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

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

Westworld. (2016). [TV series] HBO.