The AAAI-17 Workshop on
AI, Ethics, and Society
WS-17-02

Social Attitudes of AI Rebellion: A Framework

Alexandra Coman,[1] Benjamin Johnson,[1] Gordon Briggs[1], and David W. Aha[2]

[1]NRC Postdoctoral Fellow; Navy Center for Applied Research in AI; NRL (Code 5510); Washington, DC

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

{alexandra.coman.ctr.ro, benjamin.johnson.ctr, gordon.briggs.ctr, david.aha}@nrl.navy.mil

Abstract

Human attitudes of objection, protest, and rebellion have undeniable potential to bring about social benefits, from social justice to healthy balance in relationships. At times, they can even be argued to be ethically obligatory. Conversely, AI rebellion is largely seen as a dangerous, destructive prospect. With the increase of interest in collaborative human/AI environments in which synthetic agents play social roles or, at least, exhibit behavior with social and ethical implications, we believe that AI rebellion could have benefits similar to those of its counterpart in humans. We introduce a framework meant to help categorize and design Rebel Agents, discuss their social and ethical implications, and assess their potential benefits and the risks they may pose. We also present AI rebellion scenarios in two considerably different contexts (military unmanned vehicles and computational social creativity) that exemplify components of the framework.

Society, Ethics, and AI Rebellion

In human social contexts, attitudes of resistance, objection, protest, and rebellion are not necessarily destructive and antisocial; they serve a variety of fundamentally positive, constructive social functions. At a macro-societal level, protest can support social justice. At a micro level, saying "no" in a constructive way can help maintain healthy balance in personal and professional relationships (Ury, 2007). In many cases, rebellious attitudes are arguably not merely acceptable, but ethically obligatory, e.g. an engineer refusing to continue working on a project if a number of safety issues are not addressed.

In contrast, AI rebellion is generally perceived as being fundamentally destructive: not just antisocial, but anti-human, a narrative reinforced by numerous sci-fi depictions in which AI follows in the footsteps of various mythical creatures to play the part of the ominous "other". Such manifestations of rebellion are generally attributed to post-singularity AI with mysterious but decidedly dangerous inner workings.

We believe that AI attitudes of *constructive* rebellion can in many ways contribute to "maximizing the societal benefit of AI", an AI research priority expressed by Russell, Dewey, and Tegmark (2015), by enabling refusal of unethical behavior, supporting value alignment with human groups (e.g., through protest *on behalf* of humans), maintaining safety, supporting task execution correctness, enhancing social co-creativity, and providing or supporting diverse points of view.

As we will show through two scenarios and various smaller examples, such instances of AI rebellion neither require human-level intelligence or superintelligence nor involve rebelling against humanity as a whole. We are especially interested in collaborative, human-AI interaction environments, such as the long-term collaborations envisioned by Wilson, Arnold, and Scheutz (2016). In such contexts, AI rebellion has benefits comparable to those it has in human social contexts and associated risks pertaining to the maintenance of the long-term collaborations. The two scenarios that we present are drawn from the fields of (1) military unmanned vehicles and (2) computational social creativity. The first scenario is based on pre-existing work of established practical interest, while the second is largely speculative.

To facilitate this discussion, we define AI Rebel Agents and propose an initial framework for their study. A reduced version of this framework is described in (Aha and Coman, 2017).

Rebel Agents are AI agents that can develop attitudes of opposition to goals or courses of action assigned to them by other agents, or to the general behavior of other agents. These attitudes can result in resistance, objection, and/or refusal to carry out tasks, or in challenging the attitudes or behaviors of other agents. We use "rebellion" as an umbrella term covering reluctance, protest, refusal, rejection of tasks, and similar stances/behaviors.

We call an agent against which one rebels an **Interactor**. We assume that the Interactor is in a position of power in relation to the Rebel Agent; the source(s) and nature of that power can vary. The Interactor can be human or synthetic, an individual or a group.

A Rebel Agent is not intended to be permanently adversarial towards the Interactor or in a rebelling state by default. A Rebel Agent has potential for rebellion that may or may not manifest based on external and internal conditions.

An AI agent can be specifically designed to be a Rebel Agent (**rebel by design**), but rebellious behavior can also emerge unintendedly from the agent's autonomy model (**emergent rebellion**).

Our proposed framework for AI rebellion includes types of rebellion and stages of the rebellion process. The framework is applicable to both types of rebellion introduced above: (1) it can be used to guide the development and implementation of intentionally Rebel Agents, and (2) to categorize and study the rebellion potential and ramifications of emergent rebels (including their dangerous AI potential: while we argue that AI rebellion *can*, in certain instances, be positive and beneficial, we do not claim that it is *necessarily* so). The framework also (3) facilitates discussion of social and ethics-related aspects and implications of rebellion: we demonstrate this by examining dimensions of AI rebellion with strong social implications (emotion and social capital: particularly, trust) and by including ethics-related questions pertaining to the framework throughout the paper.

Our framework is meant to be generally applicable to AI agents, with no restrictions on agent architecture, paradigm, purpose, deployment context, or other factors. However, the type and features of the agent will affect how it instantiates the components of the framework.

Rebellion Awareness

For humans, saying "no" can be a difficult, but necessary action (Ury, 2007). Depending on whom we are saying "no" to, what we are refusing to do, and the greater context, this action may entail considerable emotional stress and social risk, and may require ample preparation. For existing AI agents, this social and emotional baggage pertaining to rebellion is generally not modeled.

However, any agent, natural or synthetic, that rebels in a social environment is subject to the social implications of rebellion, irrespective of whether it is "aware" of these implications and of whether these implications are relevant to its goals or motivation. Humans are inherently not just rebellious, but rebellion-aware, as evidenced even by our inclination to ascribe rebellious intentions to AI.

We define **Rebellion-Aware Agents** as agents that model rebellion (their own or that of others) and reason about its implications, such as associated social risks. Things that an agent of this type might attempt to assess include: (1) whether a human teammate is 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 as such. Both of these situations raise ethical issues.

The categories of Rebel Agents and Rebellion-Aware Agents can overlap, but are not identical. That is, some Rebellion-Aware agents may not rebel themselves, while some Rebel Agents may not be "aware" of or able to reason about the implications of their rebellion.

Naïve Rebel Agents are rebellion-unaware: they can reason about the motivating factors of rebellion (defined in the next section) and deliberate on whether to trigger rebellion or not, but do not reason about the implications, consequences, and risks of the rebellious attitudes themselves.

Conflicted Rebel Agents are rebellion-aware: they can 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 own motivating factors and the anticipated consequences of rebellion. *Ethical implications pertaining to conflicted Rebel Agents include the possibility of the agent using deceptive practices to minimize the social risk associated with its rebellion.*

AI Rebellion Framework

We propose the following framework for AI rebellion, to be expanded in future work. It includes Interactor categories, stages of rebellion (which can be interpreted in various ways, with some of them possibly intertwined or missing), and factors of rebellion.

The Interactor: Identity and Power Source(s)

We categorize Interactors based on their **relationship** to the Rebel Agent, their identity as **human or synthetic**, and their **source(s) of power** in relation to the Rebel Agent.

Based on these factors, examples of Interactors include: a human operator, a human or synthetic teammate, and a mixed human/synthetic out-group (in psychology, the terms "in-group" and "out-group" refer to social groups that a subject does and, respectively, does not identify as being part of (van Stekelenburg and Klandermans, 2010)).

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". French and Raven (1959) influentially define the following bases of power in inter-human relationships: **legitimate power** (based on perception of "legitimate right to prescribe behavior"), **reward power** (based on perceived "ability to mediate rewards"), **coercive power** (based on perceived "ability to mediate punishments"), **referent power** (based on "identification with" the individual/group in the position of po-

wer), and **expert power** (based on perception that the individual/group in the position of power has "some special knowledge or expertness"). Of course, one individual/group can exert more than one type of power over another individual/group.

We can consider a human operator to have legitimate power over its AI agent(s) by default. An AI Agent that is susceptible to positive and negative reinforcement can be subject to reward power and coercive power. Notably, power sources have clear subjective components (i.e., as seen above, one is subject to the power of another only if one perceives oneself as being subject to that power), hence they will at least partially depend on the agent's modeling of them, if any.

A Rebel Agent that is part of an in-group which also includes humans could acquire, out of solidarity, and with or without awareness of it, the status of being subject to the type(s) of influence exerted by the Interactor (e.g., a socially-influential out-group) on the humans in the in-group.

More interesting situations can occur in mixed human/synthetic multi-agent environments in which agents have different types of power over one another, e.g., a human with coercive power and an AI Rebel Agent with expert power. In this example, the human agent might be able to "punish" agents which/who do not follow commands by restricting their access to various resources, while the AI agent possesses domain expertise that is essential to the team's activity, and uses its expert knowledge to reason about whether to strictly follow commands. Allowing for such situations does not contradict our definition of the Rebel Agent/Interactor relationship: while rebellion requires the Interactor to have some type(s) of power over the Rebel Agent, this does not preclude a concomitant inverse power relationship (i.e., the Rebel Agent also having power over the Interactor). Furthermore, multiple Interactors can have power (of different types or of the same type) over the same Rebel Agent; if their commands contradict one another, obeying one Interactor may mean disobeying another.

Types of Rebellion

We introduced **rebellion by design** and **emergent rebellion** in the first section, and **conflicted** (rebellion-aware rebel) and **naïve** (rebellion-unaware rebel) agents in the second section.

We propose further rebellion types based on the following 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), **interaction initiation** (reactive and proactive), **normativity** (normative, non-normative, counter-normative), **egoism** (egoistic/altruistic), **action/inaction**, and **individual/collective action**. Several of these types are partially based on terminology used in social psychology by Wright, Taylor, and Moghaddam (1990), with modifications to the meanings of some terms.

Explicit rebellion occurs in situations in which it is clear who the Interactor is and the Rebel Agent's behavior is clearly identifiable as rebellious (e.g., refusal to conduct a task assigned by a human operator).

Implicit rebellion occurs when the Interactor is not clearly defined and/or the Rebel Agent's behavior suggests or could be interpreted as rebellion, but is not clearly expressed as such (e.g., expressing an opinion that differs from the majority's or behaving contrary to social norms).

Inward-oriented rebellion is focused on the Rebel Agent's own behavior (e.g., the agent refuses to adjust its behavior as requested by an Interactor).

Outward-oriented rebellion is focused on the Interactor's behavior, to which the Rebel Agent objects. For example, the agent might confront a human Interactor whom it identifies as mistreating another human.

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).

In **proactive** rebellion, 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.

Non-conformity is inward-oriented, proactive rebellion: refusing to adjust one's behavior in order to "fit in".

Normative rebellion consists of taking protest action within the confines of what has been explicitly allowed (e.g., questioning without disobeying).

Non-normative rebellion consists of behavior that has been neither explicitly allowed nor explicitly forbidden, but diverges from the specific commands the agent has been given.

Counter-normative rebellion consists of executing actions or pursuing goals that have been explicitly forbidden.

Individual action is rebellious action conducted by a single Rebel Agent.

Collective action occurs when multiple agents are involved in concerted rebellious action.

Rebellion is **egoistic** when the agent rebels in support of its own well-being (whatever meaning that might have to the agent) or survival.

Altruistic rebellion occurs when the agent rebels on behalf of a group or in support of a group's interests.

In many cases, egoistic and altruistic rebellion can coexist, with the agent's own values being aligned with those of human groups so that it effectively "identifies" with them. *Purely egoistical AI rebellion as well as solidarity-driven altruistic rebellion on behalf of an AI-only group could be argued to be strictly ethically prohibited, at least as long as we do not have sentient AI that can be victimized.*

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. Not executing a command *does* fall under this category, as it is outwardly perceivable rebellious behavior.

In **inaction** situations, the agent develops an internal negative attitude towards a goal, task, or another agent's behavior, but does not (yet?) manifest it outwardly.

Factors of Rebellion

In social psychology, several factors that can lead to human rebellion have been identified (van Stekelenburg and Klandermans, 2010). These include grievance, frustration, and perceived injustice. As aggrieved people do not necessarily protest, social psychology has also explored factors that determine whether a person or group who has reasons to protest will actually do so. These additional factors include efficacy ("the individual's expectation that it is possible to alter conditions or policies through protest" (van Stekelenburg and Klandermans, 2010, drawing on the work of Gamson, 1992)), social capital, access to resources, and opportunities.

Based on these insights, we distinguish between two types of rebellion factors: motivating and supporting factors. Certain factors (e.g., emotion, as will be discussed) can be in either category, depending on context.

Motivating factors provide the primary drive for rebellion. For example, human social protest may have perceived inequity as motivating factor.

Supporting factors contribute to assessing whether a rebellion episode will be triggered, and/or how it will be carried out. Naïve Rebel Agents are less likely to require supporting factors, as they do not reason about the implications and potential consequences of their rebellion.

Generally, some form of divergent access to information of the Rebel Agents and the Interactors is at the root of rebellion episodes. This information could be objective, but only partially available to a proper subset of the agents involved in the rebellion episode; or it can be 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 allowing for similar motivation models. The motivating factors of AI rebellion will hence not be general, but depend on the individual agent's architecture (including its motivation model, if any), interaction context, and purpose. The following list provides several examples of motivating factors. These factors do not need to operate in isolation: the combined action of several of them can initiate and sustain rebellion, and inform the rebellious action taken.

Ethics and safety: Rebel Agents can refuse tasks they assess as being ethically prohibited or violating safety norms (Briggs and Scheutz, 2015).

Team solidarity: 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 as well as saying "no", constructively, to one's own teammates, when necessary.

Intentionality: An agent that can assert itself convincingly encourages an intentional stance: the attitude that the agent is rational, and has beliefs, desires, and goals (Dennett, 1987). In human-robot interaction research, 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).

Ethics question: would it ever be ethically acceptable for an AI agent to protest purely for the purpose of appearing intentional, with the apparent reason of its rebellion actually being of little matter to it (similarly, in terms of mismatch between apparent and real motivation, to a child trying to assert her personality by refusing to wear an outfit chosen by her parents, not because she dislikes it, but because she has not chosen it herself)*?* This might even contravene the requirement for Rebel Agents not to be contrarian for the sake of being contrarian.

Contradictory commands from multiple Interactors: When an agent is subject to the power of more than one Interactor, there exists the possibility that obeying one of the Interactors 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; this is something that even a naïve Rebel Agent could achieve. A more complex approach could involve reasoning about consequences of rebelling against each of the Interactors, and making a decision based on trade-offs; this exemplifies conflicted rebellion.

Self-actualization: Like its human counterparts, an AI Rebel Agent could object to an assigned task that it

assesses as not playing up to its strengths or not constituting a valuable learning opportunity.

Stages of Rebellion

Pre-rebellion: This stage includes processes leading to rebellion, including development of the agent's motivation and observation and assessment of changes in the environment that are relevant to the agent's motivation. The progression towards rebellion may be reflected in the 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 (e.g., within pre-rebellion or rebellion execution) in which motivating and supporting factors of rebellion are assessed to decide whether to trigger rebellion.

Ethics questions: Should an AI agent be required to always signal to humans that it is considering rebellion, even if it does not end up rebelling? Can failure to do so be considered a type of deception? The answers might vary based on the nature, purpose, and operational context of the agent.

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? 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 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 (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 Interactor's reaction 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 can be reflected in the agent's outward behavior.

These stages of rebellion 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.

Social Dimensions of Rebellion

Social Capital, Trust, and Rebellion

Van Stekelenburg and Klandermans (2010) name social capital as a decisive factor in a human's decision to protest. In social psychology, trust is considered a component of relational social capital. Abbass et al. (2016) define trusted autonomy in a collaborative human/AI context as "a situation in which an autonomous agent willingly becomes *vulnerable* by delegating a task to itself or another autonomous agent." Previous definitions also linked trust to vulnerability (Lee and See, 2004).

The connection between rebellion and trust is inherent and multifaceted (and was previously briefly explored by us in (Johnson et al., 2016)): a moment of rebellion is inherently one of vulnerability on multiple levels. By rebelling, an agent (1) makes itself vulnerable, and (2) creates vulnerability in the system it is (or was originally) part of. Rebellion can affect trust both negatively and positively, while trust and distrust can be factors of rebellion. In fact, in situations involving only human agents, it can be argued that there are trust-related factors behind any decision to rebel. Rebellion always puts a strain on trust, but does not necessarily diminish it. Trust can increase after instances of rebellion, depending on what has caused the rebellion, and how post-rebellion was conducted by the agents involved. Whether rebellion influences trust negatively or positively also depends on what the rebelling agent was trusted to do and/or not to do (e.g., was it trusted to strictly follow commands or to behave autonomously and efficiently in new situations?). Repeated rebellion-type interactions that prove to have been justified and are positively resolved can lead to agents progressing from the former type of trust to the latter one).

Here are examples illustrating connections between rebellion and trust:

- Rebellion can diminish the trust of other agents in the Rebel Agent. For example, a human operator assigning a task to an AI agent and trusting the agent to execute that task will find their trust violated if the agent rejects the task.
- Rebellion can cause the trust of other agents in the Rebel Agent to increase. An agent that does not necessarily follow commands can gain trust in the same way in which a human expert in a field would by raising objections when appropriate.

Here are examples of trust and distrust acting as rebellion factors (assuming rebellion-aware and trust-aware agents):

- Agent A will not rebel unless it "believes" that it is trusted sufficiently to "afford" to do so. Its rebellion is not necessarily a violation of this trust. The agent could be an expert reasoning: "I know that I am trusted when it comes to my field of expertise, so if I take a different course of action than instructed, I will be trusted to have done so for good reasons."
- Agent B will not rebel unless it trusts other agents to handle the vulnerabilities that the rebellion creates both for agent B itself and for the system as a whole. In this case, trust flows in the opposite direction from the one in the previous example.
- Agent C rebels because, based on learned trust (Marsh and Dibben, 2003), it distrusts the agents it has been assigned to collaborate with, thus reasoning that it cannot entrust its vulnerabilities to them if the task is to be executed successfully, with safety maintained.
- Agent D's current level of trust in itself (e.g., in its own ability to accurately assess the current situation as warranting opposition) influences its decision to rebel.

Emotion and Rebellion

Ethics Question: Is it ethical for AI attitudes of rebellion to be in any way emotionally-charged, either in terms of their triggering mechanisms or in terms of their manifestation?

In social psychology, van Stekelenburg and Klandermans (2010) list emotion (notably anger) as a key factor of human protest. Emotion is currently studied in the following intelligent agent contexts referred to collectively as "affective computing" (Picard, 2003): (1) simulating displays of emotion, (2) acquiring and replicating models of human emotion (Gratch and Marsella, 2004), and (3) as a driving component of cognitive processes such as learning (e.g., in some approaches to intrinsically-motivated reinforcement learning (Sequeira, Melo, and Paiva, 2011)). These research directions make possible the following roles for emotion in our framework:

1. Displays of emotion being used as outward manifestations in pre-rebellion, rebellion execution, and post-rebellion.
2. Simulated emotions being used as determinant or supporting factors during rebellion deliberation (e.g., anticipatory emotions, hope and fear (Moerland, Broekens, and Jonker, 2016), experienced in pre-rebellion and used as rebellion triggers).
3. Models of the Interactor's emotional states being used to determine whether a rebellion episode would be opportune. This would, of course, raise ethical questions with regard to possible manipulation attempts.
4. Models of human teammates' emotional states being used to decide whether to rebel on their behalf, against an outside Interactor (e.g., "My human teammates appear to be under too much stress to handle this additional task. Please reconsider this assignment!") Misuse of this ability could lead to behavior reminiscent of the misguided mind-reading robot in the short story "Liar!" by Asimov (1950). General ethical implications of human emotion modeling by AI are explored by Picard (2003).

Let us further consider the ramifications of options (1) and (2) above. *Should simulated emotion ever be used as a way of interfacing with humans in rebellion contexts (e.g., for expressing rebellion (Briggs and Scheutz, 2015))? Should it be allowed to have a deeper role in the mechanisms of rebellion, serving as factor in rebellion deliberation?* In the former case, Rebel Agents could use emotionally-charged persuasion to prey on humans (e.g., HAL's pleading (Kubrick, 1968)); in the latter, Rebel Agents themselves might fall prey to emotions similarly to humans, making their rebellious behavior dangerously irrational and unpredictable. Such ethical issues are not unique to Rebel Agents, but arise from exploration of emotional/emotion-aware AI in general (emotion is considered a factor on one of the "pathways to dangerous AI" by Yampolskiy, 2016).

As a final example of emotion-fueled rebellion, we speculate that emotional contagion (Saunier and Jones, 2014) could be used to spread rebellion to multiple agents, inciting collective action, with possibly problematic implications. This raises questions such as: (a) could rebellious attitudes, through emotional contagion, "infect" agents that have no prior motivation for rebellion, or (b) could contagion lead to concerted rebellious action by agents driven by heterogeneous motivating factors (as in the case of human group protest in which the participants might be motivated by different types of grievances, but driven to protest participation by shared anger)?

Related Work

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).

Coman, Gillespie, and Muñoz-Avila (2015) proposed Rebel Agents in a more limited context: goal reasoning in interactive storytelling, meant to enhance character believability and provide a source of conflict, a key aspect of narrative in any medium. We expand and generalize their definition.

Other prior work (some of which we cited above) also falls under our definition of Rebel Agents and can be classified and analyzed according to our framework. We provide several examples here.

Gregg-Smith and Mayol-Cuevas (2015) present handheld smart tools that "refuse" to execute actions which violate task specifications. Their work exemplifies task execution correctness as a motivating factor and behavioral rebellion expression (physically resisting incorrect movements).

Briggs and Scheutz (2015) focus on the rebellion deliberation stage: they propose a general process for embodied AI agents' refusal to conduct tasks assigned to them 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) demonstrate ways in which embodied AI agents can convincingly express their reluctance to perform a task: their focus is, hence, on expressing rebellion, through verbal and nonverbal communication.

Hiatt, Harrison, and Trafton (2011) propose agents that use theory of mind to determine whether they should notify a human that he/she is deviating from expected behavior. This exemplifies outward-oriented, proactive rebellion. In another example in this category, but addressing 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.

Work on artificial moral agents (Wiegel, 2006; Kuipers, 2016) is also highly relevant to us, as the ethical permissibility of assigned tasks is one of the most significant potential factors of rebellion. Also pertinent are human-robot interaction studies on how humans respond to persuasion attempts by AI (e.g., Stock, Guerini, and Pianesi, 2016).

Disaster Relief Mission Scenario: Normative Rebellion as Safeguard

Scenario Description

For the first example scenario, consider an AI agent participating in a disaster relief mission. This agent is among a small group of autonomous agents that operate within a play calling architecture, as described by Apker, Johnson, and Humphrey (2016). This team receives commands from a centralized (human or artificial) Interactor. These commands are treated as inputs to a finite-state automaton (FSA), which is synthesized from a set of pre-defined templates. The templates, which are based on the tasks and vehicles included in the scenario, define the structure of a play. Each play structure includes the following, associated with a single command:

1. A primary behavior, associated with the command, which the agent is expected to execute.
2. A region constraining the primary behavior.
3. An environmental event that, when it is sensed, triggers the activation of a secondary behavior.
4. A region constraining the secondary behavior.
5. A description of the agent's internal state that is needed to execute the required behaviors.

Additionally, each agent's specification includes a number of contingency structures, which each include the following components:

1. A description of the agent's internal state that, when true, triggers the contingency.
2. A description of the agent's internal state that is required for it to carry out the contingency.
3. The contingency behavior, to be performed when the agent's internal state satisfies (1) and (2).
4. A region constraining the contingency behavior.

Generally speaking, each autonomous agent executes a synthesized FSA that prescribes one or more desired behaviors for each command that it may receive from the Interactor. Additionally, the FSA monitors the health of the agent and prescribes a desired response (i.e., contingency behavior) for detected faults. These contingency behaviors, when triggered, supersede the behaviors that are ascribed to the Interactor's commands, causing the AI agent to rebel.

Types of Rebellion

We can now characterize the rebellion that our Rebel Agent exhibits in this scenario when it detects a fault and activates a contingency behavior. Because the rebellion is explicitly planned for in the specification templates (i.e., the system includes the contingency structures), it can be classified as a **rebellion by design** agent. As the agent

does not model or consider the effects of its rebellion on itself or other agents, it can be classified as a **naïve** Rebel Agent.

The Interactor in this scenario, be it a human operator or another artificial system, has **legitimate power**, as the play structures are defined with commands included specifically to allow the Interactor to dictate the behavior of the agents. Despite receiving a command, the agent (when it detects a fault) **explicitly** rebels by taking a different course of action than the one that was dictated by the Interactor. This rebellion can be further classified as **reactive** and **inward-oriented**, as the received command is a specific request of the Interactor, and the Rebel Agent rebels by changing its own, internal behavior.

Furthermore, the agent exhibits **normative** rebellion, as the contingency behavior constitutes an explicitly defined and allowed form of rebellion. While there may be multiple Rebel Agents within the system, the contingency behaviors that are defined for each are executed without the contribution of any other agents, making them **individual actions**.

Finally, the rebellion exhibited by a Rebel Agent in this scenario can be characterized as either **egoistic** or **altruistic**, depending on the defined contingency behavior. For example, a rebellion episode in which the agent ignores a command in order to refuel itself (because it detects that it is low on fuel) could be considered **egoistic**, as the resulting action is for the benefit of the agent itself, but has **altruistic** implications as well, because low fuel levels can jeopardize the entire mission and even place humans in danger. On the other hand, a rebellion situation in which an airborne Rebel Agent attempts to safely land after loss of localization is immediately identifiable as **altruistic**. In this case, the agent itself would benefit more from staying in the air while attempting to regain localization; landing benefits others by maintaining safety.

Factors and Stages of Rebellion

The rebellion exhibited in this scenario is motivated by the play and contingency structures used to create the agent's FSA. More specifically, the **motivating factors** of rebellion pertain to the described internal state of the Rebel Agent that triggers the contingency behavior. On a higher level, in the low-fuel and loss-of-localization examples, safety is a **motivating factor**. In this case, there are no **supportive factors**, and the **rebellion deliberation** consists of the agent monitoring its internal state and comparing it to the description of the internal state that triggers the rebellion.

Deliberation in the **pre-rebellion** stage consists simply of monitoring one's state while following the commanded behavior.

Rebellion is **expressed** by the activation of the contingency behavior, which is triggered by the evaluation of the agent's internal state. This trigger may be based on the current state (e.g., detecting the loss of localization) or on a projection of the agent's state (e.g., the agent identifies that it does not have sufficient fuel to complete the assigned task and preemptively activates the refueling contingency). Regardless, the rebellion is expressed behaviorally, when the Rebel Agent changes its behavior. However, the system could be adjusted to require the Rebel Agent to notify the Interactor when it activates a contingency behavior; the rebellion would then be both behaviorally and verbally expressed.

The **post-rebellion** behavior of the agent is similar to its behavior during the rebellion deliberation stage. After the agent has selected a contingency behavior, it continues to monitor its internal state while it executes that behavior. If its internal state changes, it may also change the agent's behavior. For example, if the agent loses localization and rebels by attempting to land safely, it will continue to monitor its localization during the landing process; if in doing so it regains its localization, and has no other state changes, it will elect to end its rebellion and return to the commanded behavior. Likewise, after an agent rebels due to low fuel levels, it returns to its base station and refuels; once it finishes refueling, it returns to the commanded behavior it received from the Interactor. There is no post-rebellion behavior in which the consequences of rebellion are specifically targeted. The agent is naïve with regard to both trust and emotion.

Disaster Relief with Complex Deliberation

As an example of more complex rebellious behavior in a disaster relief context, consider the following variant. An autonomous transport vehicle is carrying urgently-needed food, water, and medical supplies to a remote village in an area decimated by a powerful earthquake. Unfortunately, the main route toward the transport's destination has been severely damaged: a bridge over a vital stream crossing has collapsed. The transport stops and radios back to headquarters to report the current predicament. The operator at HQ tasks the transport with attempting to ford the stream. At this point, the autonomous transport needs to consider and either accept or reject this alternate course of action.

Following the framework presented by Briggs and Scheutz (2015), this **rebellion deliberation** decision-making process should explicitly consider a number of criteria:

1. **Capability**: Is the robot properly equipped for fording rivers? Is the robot properly equipped to ford rivers of the present depth and current?

2. **Knowledge**: Does the agent know how to perform a safe ford? Does the agent know the depth and current conditions of the river?
3. **Obligation**: Is the agent issuing this directive authorized to do so? Is the agent receiving the directive obligated to obey it?
4. **Permissibility**: Is there a reason why it would be permissible to disobey this directive?

Implementing checks of this sort would require a large amount of sophisticated perceptual capabilities and integration with formal normative reasoning mechanisms. Assuming the first three criteria are upheld, a principal reason to reject this suggestion is that the river crossing is unsafe, and may incapacitate, damage, or destroy the vehicle and/or its payload.

For the first two criteria, compliance in some counterfactual cases is not ruled out (e.g., "I would cross if I knew the depth of the river."), so the agent's response may not even be interpreted as rebellious by the Interactor.

As an example of rejection based on the obligation criterion, consider if instead of the operator from HQ suggesting the attempted ford of the river, a low-level relief worker at the bridge suggested it. This worker is likely not authorized to make major alterations to the plans of the robot, as helpful as they may be.

With regard to permissibility, assuming that the robot had knowledge of its mission and the role/value of its cargo, then it is likely that such an instance of rebellion would be considered **altruistic** in addition to **egoistic**, given the consideration of the harm that would befall those in the village at the loss of the payload. However, consider a variant of this scenario in which the transport is empty (perhaps it is needed to transport wounded people from the village to better medical facilities). The robot may not itself have knowledge of these goals, having simply been instructed to drive to the village as quickly as possible. Should the robot still reject the command based on a principle of avoiding self-harm?

The case of a Rebel Agent preventing harm or damage to itself is an intriguing one, as it evokes the negative sci-fi connotations of egoistic robotic rebellion. However, it also reflects compliance to Asimov's Third Law (Asimov, 1950). There will usually be some altruistic component of self-preservation as well. Most robots are, for the time being, expensive assets, and human operators and the organizations that own the robots and employ the operators would likely be averse to instructing robotic assets to engage in needlessly self-damaging behavior.

Speculative Scenario: Creative Rebellion, Rebellious Creativity

Scenario Description

Rebellion and creativity go well together. There are numerous examples in the history of human artistic endeavor of creative artifacts that caused controversy when first introduced due to being perceived as counter-normative, only to later became critically acclaimed, popular, and influential. This holds for specific instances of art (e.g., Stravinsky's "The Rite of Spring" in 1913: the music and the ballet choreography) as well as for entire genres (e.g., at various times in musical history: jazz, rock'n'roll, rap music). Also, clashes that ultimately enhance the artifacts produced are common when it comes to humans collaborating on creative endeavors (e.g., band members, writers and editors, actors and directors).

This scenario is inspired by the research areas of (1) computational social creativity (Guckelsberger et al., 2016), (2) interactive narrative guided by drama managers (Riedl and Bulitko, 2012; Sharma et al., 2010), and (3) perpetual learning (Roberts et al., 2016). In this context, the Rebel Agent becomes a Rebel Artist or Rebel Actor. We briefly describe the three areas below.

Computational social creativity: In computational social creativity contexts, multiple agents, human and synthetic, conduct co-creative activities that generally result in artifacts, either ephemeral or persistent. Such activities include collaborative musical improvisation (Weinberg, Driscoll, and Thatcher, 2006), sketching (Davis et al., 2015), dance (Jacob and Magerko, 2015), and co-creation of narrative in a manner similar to pretend play (Magerko et al., 2014).

Guckelsberger et al. (2016) actually state: "If we want artificial agents to be taken seriously as partners in creative activities, ***we require them to challenge us***". They also point out that "co-creative and social creativity systems are only meaningful if each agent has a different perspective on a shared world, allowing them to complement each other, and for creativity to emerge from their interaction". Similarly, differences in world perception or interpretation by various agents are often at the root of attitudes of rebellion.

Guckelsberger et al. (2016) propose two types of agents in co-creative contexts: **antagonistic agents**, which impose temporary constraints on other agents to guide their creative/learning processes (with too many constraints potentially stifling creativity) and **supportive agents**, which assist one another in being creative within the established constraints.

Interactive narrative: In the intelligent systems approach to interactive narrative (Riedl and Bulitko, 2012), in which

an interactive storytelling system must respond to player actions that damage narrative coherence, AI drama managers are used to generate new narrative trajectories when user actions render the current one unusable (e.g., if the player character causes the death of a non-player character who was meant to play a crucial role later on in the story, a different non-player character may be directed by the drama manager to play that crucial role).

Perpetual learning: Roberts et al. (2016) recently proposed a type of learning, envisioned as a long-term process occurring over a variety of tasks, in which the learning agent gains increased control over its own learning process (e.g., choosing what types of tasks to learn and when to initiate or stop a learning process).

Let us assume a co-creative interactive narrative environment populated by antagonist and supportive agents, any of which can be Interactors in rebellion episodes. This environment would combine creative elements from literature, film, and emergent multi-user gaming. There are AI Rebel Actors playing parts under the guidance of a drama manager as well as human participants interacting with these actors. The drama manager could be an AI agent, but it could also be a human who is thus provided with a "film director" role.

The AI actors are perpetual learners: they learn over multiple acting experiences in different story worlds and scenarios, from various directors as well as various humans they interact with, and are not constrained to only one role over their lifetime. By learning, they improve their "acting skills" and knowledge of what constitutes believable behavior. This approach to AI actors would, to the best of our knowledge, be novel, with existing AI actors in games effectively identifying with their characters (e.g., the motivations/goals of an AI agent playing the part of a medieval knight might include "self-preservation" and "victory in battle", as opposed to "successfully portraying a medieval knight in a convincing way" – which would be the motivation/goal of a human actor in that situation). Samsonovich and Aha (2013) proposed a distinction between characters and actors in narrative-oriented goal reasoning: our proposed scenario would be a way of demonstrating that distinction in a meaningful way.

Types, Factors and Stages of Rebellion

The director (who, in this case, is a primarily **antagonistic** agent in the creative process, constraining the actors' behavior for the benefit of the emerging narrative), on observing the in-story "death" of a non-player character who was meant to later execute a task crucial to the story, attempts to re-assign that task to another character, played by a different Rebel Actor. However, this Rebel Actor, based on its prior experience and knowledge of the part, decides that executing the re-assigned task would not constitute believable behavior for its character. In this situation, the director may have **legitimate power**, while the agent itself has **expert** power (i.e., it is an "expert" in its own part)[1].

The **motivating factors** in this rebellion scenario can be character believability (if the actor refuses an action assessed as being out-of-character) and the agent's own self-actualization as an actor (if it refuses taking on an entire part assessed as being unsuitable for it or not constituting a valuable learning experience). **Supportive rebellion factors** (assuming a **conflicted** Rebel Actor) could include the agent's assessment of how much it is trusted by the director and/or how much social capital it has.

Pre-rebellion, in this case, includes the learning process that will eventually equip the actor for rebellion.

Rebellion deliberation might occur based on questions such as: "Can I handle this part?", "Will the part challenge me?", and "Am I established enough as an actor to make such demands?" The latter question would, again, be indicative of a conflicted Rebel Actor reasoning about the potential consequences of its rebellion.

A conflicted actor might also see itself as being in a valuable collaboration relationship with its director, so **post-rebellion** might consist of managing that relationship.

Reactive, **inward-oriented** rebellion might consist of refusing specific requests of the director.

Proactive, outward-oriented, altruistic rebellion can consist of confronting the director on behalf of other actors, human or synthetic, perhaps eventually leading to **collective** protest action. A script change could also be **proactively** requested of the director.

The intended audience could also be seen as a group Interactor with reward and coercive power over the Rebel Actor, which may or may not **conform** to audience expectations.

While all this would certainly constitute "rebellious creativity", how could it gain the title "creative rebellion" as well? Here is one possibility: what if the agent were to use its acquired acting skills to express rebellion to the Interactor in a convincing, emotionally-charged way? In this case, the transfer of "acting skills" used on the narrative plane to rebellion expression on the plane of interaction with the director would raise various ethical acceptability issues[2]. Ideally, a generalization of this

[1] Mutually-challenging work relationships between actors and directors sometimes produced masterpiece films, such as when director Werner Herzog and lead actor Klaus Kinski intensely disagreed over how the title part should be played in a film that would become critically acclaimed (Canby, 1977): "Aguirre, The Wrath of God" (1972).

[2] A fictional, "dangerous/victimized AI" portrayal of such a transfer of skills (as well as other aspects of AI rebellion) can be seen in the TV show "Westworld" (2016).

transfer, i.e. transfer of *positive* rebellious drive and behavior from fictional to real scenarios, could support value alignment in non-entertainment interaction environments (the connection between narrative intelligence and value alignment is explored by Riedl and Harrison, 2016).

Conclusion

We have introduced Rebel Agents and a general framework for AI rebellion that can be used to implement new intentionally Rebel Agents, analyze preexisting agents for rebellion potential with its various ramifications, and frame conversations about the socio-ethical implications, benefits, and risks of AI rebellion.

References

2001: A Space Odyssey. 1968. [video] United Kingdom/United States of America: Stanley Kubrick.

Abbass, H.A.; Petraki, E.; Merrick, K.; Harvey, J.; and Barlow, M. 2016. Trusted Autonomy and Cognitive Cyber Symbiosis: Open Challenges. *Cognitive Computation* 8(3): 385–408.

Aguirre, The Wrath of God. 1972. [video] West Germany: Werner Herzog.

Aha, D.W., and Coman, A. 2017. AI Rebellion: Changing the Narrative. In *Proceedings of Thirty-First AAAI Conference on Artificial Intelligence*. *To appear*.

Apker, T.; Johnson, B.; and Humphrey, L. 2016. LTL Templates for Play-Calling Supervisory Control. In *Proceedings of AIAA Science and Technology Forum Exposition*.

Asimov, I., 2004. *I, Robot* [1950] New York: Bantam Dell.

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 *Proceedings of the International Conference on Social Robotics*, 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.

Canby, V. 1977. “Aguirre, the Wrath of God” Haunting Film by Herzog. The New York Times.

Cialdini, R.B., and Goldstein, N.J., 2004. Social Influence: Compliance and Conformity. *Annual Review of Psychology* 55: 591-621.

Coman, A.; Gillespie, K., and Muñoz Avila, H. 2015. Case-based Local and Global Percept Processing for Rebel Agents. In Kendall-Morwick, J. (ed.) ICCBR (Workshops), volume 1520 of *Proceedings of the CEUR Workshop*, 23–32. CEUR-WS.org.

Davis, N.; Hsiao, C.P.; Popova, Y.; and Magerko, B. 2015. An Enactive Model of Creativity for Computational Collaboration and Co-creation. In Zagalo, N. and Branco, P. (eds.) *Creativity in the Digital Age*, 109-133. London: Springer.

Dennett, D.C. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.

French, J.R.P, and Raven, B. 1959. The Bases of Social Power. In Cartwright, D. (ed.). *Classics of Organization Theory*, 311-320.

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

Gratch, J., and Marsella, S., 2004. A Domain-Independent Framework for Modeling Emotion. *Cognitive Systems Research* 5(4): 269-306.

Gratch, J.; Nazari, Z.; and Johnson, E. 2016. The Misrepresentation Game: How to Win at Negotiation While Seeming Like a Nice Guy. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, 728-737. Singapore: ACM Press.

Gregg-Smith, A., and Mayol-Cuevas, W.W. 2015. The Design and Evaluation of a Cooperative Handheld Robot. In *Proceedings of the International Conference on Robotics and Automation*, 1968-1975. Seattle, WA: IEEE Press.

Guckelsberger, C.; Salge, C.; Saunders, R.; and Colton, S. 2016. Supportive and Antagonistic Behaviour in Distributed Computational Creativity via Coupled Empowerment Maximisation. In *Proceedings of the Seventh International Conference on Computational Creativity*.

Heckhausen, J.E, Heckhausen, H.E. 2010. *Motivation and Action*. Cambridge: 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 International Joint Conference on Artificial Intelligence*.

Jacob, M., and Magerko, B. 2015. Interaction-based Authoring for Scalable Co-creative Agents. In *Proceedings of the Sixth International Conference on Computational Creativity*, 236–243.

Johnson, B.; Floyd, M.W.; Coman, A.; Wilson, M.A.; and Aha, D.W. 2016. Goal Reasoning and Trusted Autonomy. *Foundations of Trusted Autonomy*. *To Appear*.

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

Kuipers, B. 2016. Human-Like Morality and Ethics for Robots. In In Walsh, T. (ed.) *AI, Ethics, and Society: Papers from the AAAI Workshop* (Technical Report WS-16-02). Phoenix, AZ: AAAI Press.

Lee, J.D., and See, K.A. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46(1): 50-80.

Magerko, B.; Permar, J.; Jacob, M.; Comerford, M.; and Smith, J. 2014. An Overview of Computational Co-creative Pretend Play with a Human. In *Proceedings of the First Workshop on Playful Virtual Characters, 14th Conference on Intelligent Virtual Agents*.

Marsh S., and Dibben, M. R. 2003. The Role of Trust in Information Science and Technology. *Annual Review of Information Science and Technology* 37(1): 465–98.

Moerland, T.; Broekens, J.; and Jonker, C. 2016. Fear and Hope Emerge from Anticipation in Model-based Reinforcement Learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

Molineaux, M., and Aha, D.W. 2015. Continuous Explanation Generation in a Multi-Agent Domain. In *Proceedings of the Third*

Conference on Advances in Cognitive Systems. Atlanta, GA: Cognitive Systems Foundation.

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

Riedl, M.O., and Bulitko, V. 2012. Interactive Narrative: An Intelligent Systems Approach. *AI Magazine* 34(1): 67.

Riedl, M.O., Harrison, B. 2016. Using Stories to Teach Human Values to Artificial Agents. In Walsh, T. (ed.) *AI, Ethics, and Society: Papers from the AAAI Workshop* (Technical Report WS-16-02). Phoenix, AZ: AAAI Press.

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 Humphrey, L.; Topcu, U.; Singh, S.; Miller, C.; and Vardi, M. (eds.) *Cross-Disciplinary Challenges for Autonomous Systems: Papers from the AAAI Fall Symposium (Tech. Rep. FS-16-04)*. Arlington, VA: AAAI Press.

Russell, S.J.; Dewey, D.; and Tegmark, M. 2015. Research Priorities for Robust and Beneficial Artificial Intelligence. *AI Magazine* 36(4): 105-114.

Samsonovich, A.V., and Aha, D.W. 2013. Character Oriented Narrative Goal Reasoning in Autonomous Actors. In *Goal Reasoning: Papers from the ACS Workshop (Technical Report CS-TR-5029)*, 166–181 College Park, MD: University of Maryland, Department of Computer Science.

Saunier, J., and Jones, H. 2014. Mixed Agent/Social Dynamics for Emotion Computation. In *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems*, 645-652. International Foundation for Autonomous Agents and Multiagent Systems.

Sequeira, P.; Melo, F.S.; and Paiva, A. 2011. Emotion-Based Intrinsic Motivation for Reinforcement Learning Agents. In *Proceedings of the Fourth International Conference on Affective Computing and Intelligent Interaction*, 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.

Stekelenburg, J. van, and Klandermans, B. 2010. The Social Psychology of Protest. *Sociopedia.isa*.

Stock, O.; Guerini, M.; and Pianesi, F. 2016. Ethical Dilemmas for Adaptive Persuasion Systems. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 4157–4161. AAAI Press.

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. In *Proceedings of Fifteenth International Conference on Automated Planning and Scheduling*, 161-170. Monterey, CA: AAAI Press.

Vattam, S.; Klenk, M.; Molineaux, M.; and Aha, D.W. 2013. Breadth of Approaches to Goal Reasoning: A Research Survey. In Aha, D.W.; Cox M.T.; and Muñoz-Avila, H. (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. In *Proceedings of the Eighth International Conference on Social Robotics*, 684-693. Paris: Springer

Weinberg, G.; Driscoll, S.; and Thatcher, T. 2006. Jam'aa - A Middle Eastern Percussion Ensemble for Human and Robotic Players. In *Proceedings of the International Computer Music Conference*, 464–467.

Westworld. 2016. [TV series] HBO.

Wiegel, V. 2006. Building Blocks for Artificial Moral Agents. In Allen, C.; Wallach, W.; and Brady, M. (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 Walsh, T. (ed.) *AI, Ethics, and Society: Papers from the AAAI Workshop* (Technical Report WS-16-02). Phoenix, AZ: 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): 229–250.

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.

Yampolskiy, R.V. 2016. Taxonomy of Pathways to Dangerous Artificial Intelligence. In Walsh, T. (ed.) *AI, Ethics, and Society: Papers from the AAAI Workshop* (Technical Report WS-16-02). Phoenix, AZ: AAAI Press.