In *Proceedings of the ASIMOV workshop as part of the International Conference on Social Robotics* (ICSR 2021), Singapore, Singapore, November 2021

Intelligent Disobedience and AI Rebel Agents in Assistive Robotics

Reuth Mirsky[1,2] and Peter Stone[1,3]

[1] The University of Texas at Austin, USA
{reuth,pstone}@cs.utexas.edu
[2] Bar Ilan University, Israel
[3] Sony AI

Abstract. With the increasing integration of service robots into assistive technologies, there is a need to reason about the boundaries and scope of these robots' autonomy, such as when they should merely react to their environment, when they should make proactive decisions, and when they should override commands. In most existing research, the definition of a "good" assistive robot is one that is compliant with respect to the commands it is given. Two recent papers challenge this perspective, and describe scenarios where a system might choose to rebel against a command or disobey its handler due to a deep understanding of the handler's intentions. This paper provides a comparative discussion about these two papers and how they together create a more comprehensive framework for assistive robots that can override commands.

Keywords: Assistive Robots · Multiagent Systems · Human-Robot Interactions · Autonomous Agents · Intelligent Disobedience · AI Rebel Agents.

1 Background

In most existing research on collaborative robots and agents, the definition of a "good" robot is one that is compliant with respect to the commands it is given, and that works in a predictable manner under the consent of the human it serves [5, 11]. However, as exemplified in Issac Asimov's Second law of robotics [2], this compliance is not always desired, when it might interfere with the safety of the human:

> **First Law:** A robot may not injure a human being or, through inaction, allow a human being to come to harm.
> **Second Law:** A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
> **Third Law:** A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

When discussing these laws in the context of real world domains, the second part of the Second Law (the word "except" and onward) is much richer than

simply protecting the safety of the human. For example, there are cases where the desirable behavior of the robot is not to comply with a command, even if it does not harm a human but rather the robot itself, such as the case of a controlled unmanned aerial vehicle which is steered into a wall. This case is not covered by Asimov's laws, because the safety of the robot (Third Law) is prioritized below obeying the orders given by a human (Second Law). However, many modern autonomous systems operate within the limits of predefined safety restrictions that protect both the human and the robot [1, 8, 9].

There are additional situations in which a robot should not obey a command, especially in the case of assistive robots, where the robot will need to reason about cases in which the "right" thing to do is the opposite from the instruction given by the handler. For example, a semi-autonomous wheelchair crossing a congested road should choose to brake or slow down, even if there is no immediate danger in crossing and the human driver keeps pressing the gas. This behavior is in conflict with the second part of the Second Law, even though safety is not an immediate issue. Similarly, an intelligent cane should direct a person to take an alternative route if the route the person chose is blocked. In this scenario, there is no safety issue and still an intelligent robot will be more useful if it acts differently than ordered. Two recent papers discuss these situations and others where a system might choose to rebel against a command or disobey its handler due to a deep understanding of the handler's intentions [4, 10].

Coman and Aha [4] presented the concept of **AI Rebel agents**, which are artificially intelligent agents that can refuse assigned goals and plans, or oppose the behavior or attitudes of other agents. This umbrella term can be used to describe a wide variety of artificial agent behaviors, including intentional and unintentional disobedience, a shift from the desired outcome due to misunderstanding of a command, "ethical nudges" [3], and critiquing. Coman and Aha present different dimensions of rebellious robots, such as whether the robot intends to rebel and who initiates the interaction. Next, the paper delineates the stages of a rebellious agent: pre-rebellion, rebellion deliberation, rebellion execution, and post rebellion [4]. Three different scenarios are used to exemplify different rebellion types: a furniture-moving robot, a personal assistant, and a hiring committee observer.

Recently, another paper took a closer look into the notion of ignoring or disobeying a command specifically in the context of assistive robots. Mirsky and Stone [10] (the authors of this paper) presented the seeing-eye robot grand challenge as a guiding use-case that will help imagine how to design an autonomous care system that is able to make decisions as a knowledgeable extension of its handler. The main question we considered there was whether we can design and build a service robot that can replace or surpass the functionalities of a seeing-eye dog [10]. An important function that such a robot should have is the ability to **intelligently disobey**, which consists of five stages: identifying global objectives, identifying local objectives, plan recognition, consistency check, and mediation.

2 Discussion and Open Challenges

In this section we provide a comparative evaluation of the two papers discussed above. First, the granularity of the inspected behavior differs between these two frameworks: the AI Rebel agents article discusses an act of rebellion from a more abstract point of view. Compared to the intelligent disobedience process, it does not provide details about the types of objectives a robot must consider in its pre-rebellion stage, but it does consider post-rebellion as an inherent component of the rebellion process. A more detailed examination of the different stages of the two frameworks is shown in Figure 1. As seen in the figure, pre-rebellion consists of several different abilities that the robot should have to reason about when to rebel / disobey: identifying the global objectives of its environment and handler, which also encapsulate the First Law safety constraints; identifying the local objectives of the handler, which are time- and context-specific and are tightly related to the value alignment problem; and recognizing the plan he handler which to execute. Rebellion deliberation also requires plan recognition as a way to reason about *how* the handler wishes to reach a goal, as well as the planning and verification of potential rebel acts. Finally, the AI Rebel agent framework discussed the stages of rebellion execution and post-rebellion, which have little overlap with the intelligent disobedience framework. This gap does not mean that the design of an intelligently disobedient robot should not include these stages – for example, if a seeing-eye robot decides to take an alternative route to the one proposed by the handler, it should reason about how to execute this route change (and then execute it) in a way that is explainable to- and acceptable by- the handler.

A second difference between the two frameworks is the cases they cover – generally, AI Rebel agents is more general. For example, it encompasses cases where a robot might disobey unintentionally, due to a misunderstanding of the commands given by the human or by under-specified objectives. There can be an even more subtle conflict between an instruction given by a human, and the desired outcome due to an imperfect instruction. These conflicts are examples of *the value alignment problem* [6].

Second, in the seeing-eye robot example the robot is meant to serve as an extension for its handler's desires, while an AI Rebel agent might decide to rebel due to egoistic motives. These motives can partially or indirectly overlap with the handler's (e.g., taking a detour to get fuel so it can work longer or to learn about a newly opened store the handler might wish to visit in the near future) or they can be adversarial – which is completely unwanted in the case of the seeing-eye robot. A third possibility is that the robot's motives can be completely unrelated to the handler's motives (e.g., leveraging the robot's sensors to gather research data), which might be a desired ability of assistive robots which are deployed in the real-world, but it does not overlap with the goals of the handler of the robot. Freedman and Fukunaga [7] discuss these three cases (assistive, independent, and adversarial) in the context of classical planners. While all of these cases are types of AI Rebel agents, only the first can be considered a behavior of intelligent disobedience in assistive robots.

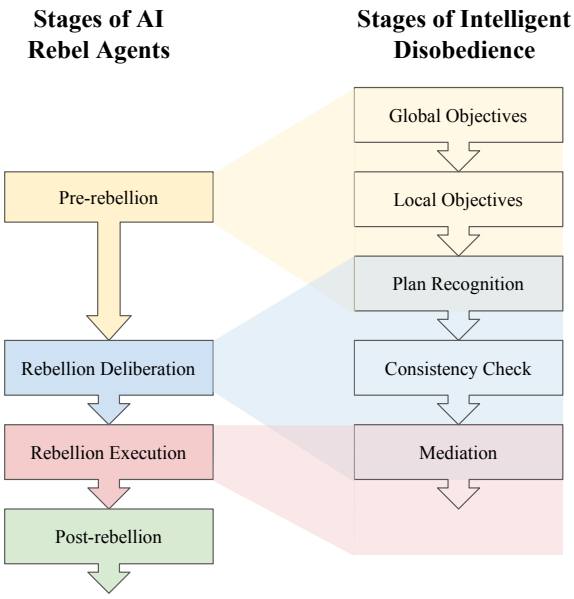

Fig. 1. Illustration of the different stages of AI rebel agents (left) and intelligent disobedience (right). As intelligent disobedience is a more specific process, colors are used to project how AI rebel agents stages can be decomposed into intelligent disobedience.

Summarizing the comparison of the two frameworks presented above, an intelligently disobedient robot can be viewed as a refinement of AI rebel agents for assistive robots. Specifically, according to the attributes suggested by Coman and Aha [4], intelligent disobedience is a specific act of rebellion in which the robot is altruistic, takes a proactive role, with an explicit design intentionality to enable the robot to disobey when needed.

3 Conclusion

As assistive robots become more prevalent as care systems, their ability to rebel – intentionally and unintentionally – must be considered in their design. Issac Asimov's laws of robotics illustrated a first, though incomplete, specification of such rebellious behavior. Modern AI and robotics research recently proposed to tackle the challenge of command override in a more structured way. The AI Rebel agents provides a general overview of the stages of a rebellion. The intelligent disobedience framework provides a first step to intentionally design collaborative robots that can rebel when such a rebellion aligns with the objectives of the human. There is still much more to be done even within the specific process of intelligent disobedience, but it is a crucial process to consider when designing assistive robots.

References

1. Agarwal, P., Deshpande, A.D.: Impedance and force-field control of the index finger module of a hand exoskeleton for rehabilitation. In: 2015 IEEE International

Conference on Rehabilitation Robotics (ICORR). pp. 85–90. IEEE (2015)
2. Asimov, I.: I, robot. Spectra (2004)
3. Borenstein, J., Arkin, R.: Robotic nudges: the ethics of engineering a more socially just human being. Science and engineering ethics **22**(1), 31–46 (2016)
4. Coman, A., Aha, D.W.: Ai rebel agents. AI Magazine **39**(3), 16–26 (2018)
5. Dragan, A.D., Lee, K.C., Srinivasa, S.S.: Legibility and predictability of robot motion. In: 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI). pp. 301–308. IEEE (2013)
6. Fisac, J.F., Gates, M.A., Hamrick, J.B., Liu, C., Hadfield-Menell, D., Palaniappan, M., Malik, D., Sastry, S.S., Griffiths, T.L., Dragan, A.D.: Pragmatic-pedagogic value alignment. In: Robotics Research, pp. 49–57. Springer (2020)
7. Freedman, R.G., Fukunaga, A.: Integration of planning with plan recognition using classical planners. In: 2015 AAAI Fall Symposium Series (2015)
8. Kapoor, A., Li, M., Taylor, R.H.: Constrained control for surgical assistant robots. In: ICRA. pp. 231–236 (2006)
9. Lankenau, A., Meyer, O., Krieg-Bruckner, B.: Safety in robotics: The bremen autonomous wheelchair. In: AMC'98-Coimbra. 1998 5th International Workshop on Advanced Motion Control. Proceedings (Cat. No. 98TH8354). pp. 524–529. IEEE (1998)
10. Mirsky, R., Stone, P.: The seeing-eye robot grand challenge: Rethinking automated care. In: Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021) (May 2021)
11. Sarathy, V., Arnold, T., Scheutz, M.: When exceptions are the norm: Exploring the role of consent in hri. ACM Transactions on Human-Robot Interaction (THRI) **8**(3), 1–21 (2019)