Only Those Who Can Obey Can Disobey: the Intentional Implications of Artificial Agent Disobedience

Thomas Arnold
Human-Robot Interaction Laboratory
Tufts University
Medford, MA, USA
thomas.arnold@tufts.edu

Gordon Briggs
Navy Center for Applied Research In Artificial Intelligence
U.S. Naval Research Laboratory
Washington, DC, USA
gordon.briggs@nrl.navy.mil

Matthias Scheutz
Human-Robot Interaction Laboratory
Tufts University
Medford, MA, USA
matthias.scheutz@tufts.edu

ABSTRACT

Recent attention has been brought to robots that "disobey" or so-called "rebel" agents that might reject commands. However, any discussion of autonomous agents that "disobey" risks engaging in a potentially hazardous conflation of simply non-conforming behavior with true disobedience. The goal of this paper is to articulate a sense of what constitutes *desirable* and *true* disobedience from autonomous systems. To do this, we begin by discussing what it is not. First, we attempt to disentangle figurative uses of the term "disobedience" from those connotative of deeper senses of agency. We then situate *true* disobedience as being committed by an agent through an action that presupposes some understanding of the violated instruction or command.

KEYWORDS

machine ethics, robot obedience/disobedience, autonomy

ACM Reference Format:
Thomas Arnold, Gordon Briggs, and Matthias Scheutz. 2022. Only Those Who Can Obey Can Disobey: the Intentional Implications of Artificial Agent Disobedience. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Auckland, New Zealand, May 9–13, 2022*, IFAAMAS, 7 pages.

1 INTRODUCTION

Robots are supposed to do what they are told. To many, the idea of a robot or artificially intelligent (AI) agent not doing exactly what is commanded by a human user raises alarming concerns: even examples of simple agents defying commands feature as a harbinger of doomsday scenarios of uncontrollable "superintelligences" [4] or a robot uprising. While the likelihood and feasibility of these speculative scenarios are debatable, their undesirability is not. More concretely, there are various examples of current, real-world, autonomous systems acting in dangerous contradiction to well-accepted human regulations and norms. For instance, some current autonomous car prototypes have been reported as "disobeying" stop signs [22], an action that would likely lead to grave human harm if not corrected. Thus, whether the speculative robot apocalypse from science fiction or the more mundane autonomous car violating traffic laws, there is a clear space of behaviors exhibiting *undesirable* "disobedience", posing dangers that must be avoided.

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Auckland, New Zealand.

Are all instances of "disobedience" by autonomous agents undesirable? Recent attention has been brought to robots that "disobey" [6] or so-called "rebel" agents [7] that might reject commands [5] for good reasons – these may be desirable instances of "disobedience." Yet, the notion of "desirable disobedience" is not new. In his fiction, Isaac Asimov posited the three laws of robotics, which specify that a robot must obey orders given by humans (second law) except when they would lead to human harm (contra first law) [3]. Another example of robot "disobedience" in the service of avoiding human harm can be seen in the domain of "seeing eye" robots [17, 23], where an assistive robot may steer its human handler away from hazards in an environment that the handler cannot perceive. The example of a "seeing eye" robot is an intuitive case where a human interaction partner may issue a command that is not aligned with his or her own overarching goals (i.e., safety). Even the most cautious skeptics of autonomous systems would likely concede that there is a space of agent behavior that could be considered *desirable* "disobedience."

Because robot and AI "disobedience" straddles the divide between desirable and undesirable behaviors, AI researchers have often approached the question of "disobedience" from the general standpoint of ensuring that any robot behaviors conform and/or are aligned with what humans consider desirable [21]. However, we contend that this approach is insufficient. In part, we argue that this is due to a potentially hazardous conflation of simply non-conforming behavior with true disobedience. Ask yourself: are all instances of robot or AI "disobedience" actually instances of disobeying something? Consider again the example of the Tesla autonomous car [22] that reportedly "disobeys" stop signs. It would seem incorrect to say that an autonomous car with computer vision algorithms that fail to correctly recognize a stop sign, or an autonomous car that has no knowledge of the difference between a complete and rolling stop, "disobeys" in a strong sense of the term. An autonomous agent with these sorts of limitations does not "disobey" in the same sense that a human-like agent would. To say that any AI or robot whose behavior does not conform to commands or norms "disobeys" may result in an over-ascription of agency and/or deliberative capability, an outcome not without its own unique set of dangers [20]. This slippage is especially important to correct given media incentives for clickbait and alarmist headlines, a propensity to magnify something like a sensor malfunction (e.g. one not picking up the lip of a doorway) as a robot breaking away for freedom [1].

The goal of this paper is thus to articulate a sense of what constitutes *desirable disobedience* from autonomous systems. To do

this, we begin by discussing what it is not. First, we attempt to disentangle figurative uses of the term "disobedience" from those connotative of deeper senses of agency. We then situate *genuine* or *true* disobedience as being committed by an agent through an action that presupposes some understanding of the violated instruction or command. For interactions within instructional contexts, we argue that true disobedience (of the instruction) depends on a broader, more important form of obedience. For the sake of design clarity and future research, it is important to delineate what kind of "obedience" ascriptions disobedience relies upon, so that accountable and transparent systems can be more squarely pursued. A robot that says "no" for a reason it can express (and justify) and rely upon is one whose obedience and disobedience can be more accessible, correctable, and functional.

2 SENSES OF "DISOBEDIENCE"

When we hear people say that something is "rebellious" or "disobedient", we can have a general understanding of what they mean. These descriptions can be applied to all manner of entities, ranging from pieces of equipment, animals, children, and adults. Yet, the implications we draw from hearing that someone's television is being "rebellious" are considerably different from the ones we draw from hearing that a person is being "rebellious". Likewise, "disobedience" has a similar range of contextually-driven connotations. Therefore, when we apply the terms "rebellion" or "disobedience" to an AI or other autonomous system, we must take care and be clear on what we precisely mean.

On one end, there is a sense of *figurative* "disobedience," in which the entity is simply behaving in a way that is unintended or undesired by a supervisory agent, but is correctly understood by the user or supervisor of being the sort of entity incapable of deciding to do otherwise. For example, we understand that a television that is failing to respond correctly to input commands from a remote control can be at best described as "disobedient" in a figurative way. On the other end, there is a sense of *true disobedience*, in which the entity is behaving in a way that is unintended or undesired by a supervisory agent, while also correctly understood by the supervisor as explicitly and deliberately deciding to do so. For example, a soldier refusing to obey an order given by a superior, citing potential violations of the rules of engagement or laws of war, would be understood as being disobedient in this truer sense.

However, this dichotomy still elides a significant amount of nuance. While obedience and disobedience may not require full agency in the sense of Moor [18], there are admittedly degrees of applicability for terms like "obedience" and "disobedience." For example, an unruly dog may not be disobedient in the same sense as a person, but we could say it is being disobedient in a sense more closely resembling true, rather than figurative, disobedience (though we will return to this point below). Furthermore, the cognitive complexity necessary for an agent to exhibit true disobedience would also likely entail a variety of contexts in which undesired behavior is not necessarily indicative of disobedience. For example, it would seem inappropriate to describe learning agents prior to or undergoing training as being disobedient. In other words, it would be less appropriate to describe an untrained dog that disregards a command as disobedient than to do so for a trained one that does the same. While a comprehensive treatment of these nuances is outside the scope of this paper, it is important to note that the degree to which the term "obedience" or "disobedience" can be applied to an autonomous agent's behavior in the non-figurative sense is directly tied with how the agent ostensibly produces these behaviors.

3 RECENT APPROACHES TO ROBOT "DISOBEDIENCE"

Recently, researchers have used partially observable Markov decision processes (POMDPs) to formalize the notion of robot disobedience as corrective realignment of a specific human interactant's specific commands and larger aims [16]. In this work, Milli et al. (2017) investigate a trade-off between robot obedience and robot autonomy: "A blindly obedient [robot] R is limited by [a human] H's decision making ability. However, if R follows a type of [inverse reinforcement learning] IRL policy, then R is guaranteed a positive advantage when H is not rational." In other words, for the robot to be able to do better when a human is not rational, it needs to be disobedient.

While this seems to be an example of true disobedience at first glance, the relationship between autonomy and obedience (and disobedience) makes that murkier. For starters, systems that are trained to simply follow policies and/or learn rewards of their instruction givers from interactions with them (e.g., through IRL) will fail to appreciate norms instruction givers might be following. They can fail to register the difference these norms represent between the instruction giver's own preferences or those of the larger society (irrespective of the instruction giver's preferences). As a result, they might view an instructor as less rational and increase their propensity to not carry out that instructor's commands, but without a specific explanation of why a particular command was not followed. Without a represented choice, or a represented reason, or – most fundamentally – some rudimentary understanding that a command/instruction *is* a command/instruction, there is not disobedience but only nonconformity. And that lack of conformity can be described just as well as "dysfunction" if it violates enough global priorities/programming as it can "disobedience". That is, only context-constrained reasons for refusals participate enough in the concept of "obedience" to be called disobedience in the true sense, rather than randomness, breakdown, and other events corresponding to figurative "disobedience." In sum, mere nonconformity in behavior does not mean something is being disobeyed.

Therefore, we would argue that robots that simply execute learned policies instead of weighing the applicability and tradeoffs among normative principles in a given context are neither obedient nor disobedient. Obedience and disobedience help define one another as ascriptions of an agent's capacities, not just external labels for behavior. Obedience thus requires the capability to disobey, which, in turn, requires an understanding of the possible norm violations implicated by particular actions, i.e., obedience and disobedience are relative to principles. A policy-based robot (regardless of how it learned the policy) is no different from a dish washer that starts the cleaning cycle when the right knob is turned. A norm-obeying robot, on the other hand, may have symbolically rendered knowledge, and (we will argue) *explicit* knowledge of normative principles that underwrite its actions. And if given a command, it can determine

what principles are implicated and then make a decision based on those principles and the ones that might potentially be violated, to choose which principle to suspend and which to uphold (e.g., see [12] for such an approach).

It should also be noted that viewing robot obedience and disobedience as a dyadic problem, that is, considering the actions of a robot relative to the commands and aims of a single human user, has significant limitations. Many instances of disobedience are considered desirable not because a rejected command violates the issuer's true intent, but rather because the command contravenes larger legal or moral principles, regardless of the intent of the command-issuing agent. For example, ethical reasoning mechanisms have been proposed for hypothetical autonomous military systems to ensure that lethal force is deployed only when authorized by the rules of engagement and laws of war [2].

What this example indicates is a need to distinguish and detail the function of disobedience in contrast to the broader set of non-conforming or unexpected behavior on the part of an artificial agent. If there is a practical demand for a robot to explain its action, one that bears on other people acting in response, then disobedience is implicated in norms whose violation could threaten more than a robot's direct task. If there are norms or principles to which a robot's action is adhering, what are they? The ambiguity of divergent action, and the ascriptions it can incur, can blur important lines of accountability. What commitments does a robot's implicit or explicit "No" really make?

4 WHAT DOES TRUE DISOBEDIENCE ENTAIL? LAYERS OF INTENTIONALITY AND INTERACTION

4.1 Understanding Instruction as Possible Action

Robots are used to perform tasks for which they have goals, implicit or explicit, regardless of how these goals made it into the system: through explicit instruction or the user's selection of one of multiple pre-defined options, or through learned policies for a given reward function. Performing a task then means for the robot to execute a sequence of actions that, if all goes well, will lead to the desired outcome. Actions taken by the robot will, in general, depend on the state of the environment, including the robot and other task-relevant objects and agents, and might thus change due to unforeseen events (e.g., the robot's effectors breaking) or actions by other agents (e.g., a human giving the robot a command that interferes with its goal). In such cases, the robot will attempt corrective actions to get back on track. In a policy-based system the best action under the current policy will be executed, while in a planning-based system the current plan will be abandoned and replanning will be triggered to determine the best course of action. Regardless of how a robot's response to circumstances that interfere with its task performance is initiated, changes in behavior to make progress towards the overall task goal per se have nothing to do with disobedience, even when these changes are the robot's reaction to a human instruction (e.g., the robot taking time to parse the speech that interrupts it).

Understanding an instruction functionally, wherein a system identifies it as a directive, is a level of intentionality that sets disobedience apart from merely failing to meet an instructed constraint. A sensory lack or a clumsy execution may prevent the system from following an instruction, but that is not at all a disobedient result. The oblivious or incompetent system merely fails at what it would do, instead of choosing failure itself.

4.2 Capability of Obedience

Thinking of robots in terms of obedience can heighten their association with animals (especially service animals or pets). "Obedience school", especially for dogs, is a common practice for getting one's pet to behave in conformity with instruction. It is worth noting carefully, however, that in that context the opposite of an obedient animal is not a disobedient one. It is an untrained one. If it does not learn some established connection between an explicit command and the expected action it is impervious to instruction: not just untrained but perhaps untrainable. This distinction, to the degree it matches one's intuitions, reflects the dependency that disobedience has on the capability to obey. That capability is more firmly assumed if there is proven performance thereof, if the agent has been discerning and reliable in obeying instructions before. If there are two dogs being commanded to sit while a bunny runs into each of their fields of vision, it is the trained one – the one that knows to assume the position that satisfies "Sit!" and under ordinary circumstances does so without hesitation – that one would ostensibly call disobedient by running away. The other dog, one that has never heard "Sit!" before, is not disobeying anything. There is nothing in its experience or learned capabilities to defy.

In child psychology, the dependence of disobedience on obedience takes on even more intentional dimensions. Seen in both everyday instruction and scholarly research, the lines between intentional and accidental action are continually negotiated as children develop the ability to interpret the intentions of other agents as well as classify, if not justify, their own actions [8]. Actions that are "unexpected" or "wrong" can be accidents, and children learn to apply degrees of intentionality before applying blame to agents, what they "knew full well" before doing [11]. The ability to obey, an established understanding and execution of its goal, informs when an agent "knew" what it was doing relative to what was expected or demanded. The ongoing moral education of children will refine differences between justified reasons and bad excuses, but throughout there will be intentional ascriptions of what the agent "knew enough" to do (hence knew enough to be held responsible for doing wrong). As one considers artificial agents and obedience, in fact, it is suggestive to look at how children interpret deviant or unexpected robotic action. Lemaignan et al [14] show that what they set up as "disobedient" behavior does not often get interpreted as such by their child participants, to the degree there is less ascription of intentionality to the robots – they already have a sense that such a behavior is not disobedience unless there is some intentionality that defines the action.

The capability of obedience is part of the implicit intentionality connecting instruction to an agent. If an agent has never obeyed a specific instruction, nor shown a response that reflects an inbuilt relationship between instruction and executable action, on

what basis would one relate its actions to the instruction at all? The perceptual proximity of an instruction does not anchor it *qua* instruction toward the agent. I may yell at a dolphin leaping close to me in the waves to come give me a ride to Australia, but what would that have to do with defining its ensuing behavior?

4.3 Reason, Purpose, or Commitment for Acting Against Instruction

The capability of obedience in a system establishes that a system is matching its action with the content of a command, not just in coincidental conformity with a command. The difference hangs on how the instruction brings about that action based on the system's design. An non-conforming action lacking any intentional relation to an instruction may be indistinguishable behaviorally from one emerging from an instruction being outweighed, overridden, or rejected on some set of terms. Disobedience, by extension, does not represent a pure severance from instruction but a more complicated relationship with it. Disobedience is not a inability to take instruction and obey. On the contrary, it is a conditional rejection of instruction, one that invites a query into what conditions explain such rejection and subsequent disobedience.

One can consider as an illustration two policy-based systems that are seeking to maximize a reward in each state they find themselves occupying. An action from each that conflicts with an instruction may have two different explanations – in the one case the system knows certain features of the state space better than the instructor, and is selecting better actions based on that knowledge. Its action just happens to conflict with an instruction that it is not incorporating into its decision. The other system simply has an estimate of the instructor's rationality that justifies partial divergence from what they instruct. Each system can be seeking to maximize reward, but they start from different points of training. If the latter system learns more about the state space than the former, it will converge in action and have its own training to depend upon – the instructor need have no further effect on the process. Disobedience need not apply to either system on an intentional level, though divergent actions from each are still observed.

In a planning system with explicit reasons (or purposes, or commitments), however, disobedience retains its sense as the deliberate rejection of an obedient action. There is a condition that meets a standard for acting against instruction, a reason that justifies taking the disobedient course.

Without such layers of reason, purpose, and commitment, one is left with other ways that work better to describe a system: errant, untrained, impervious, oblivious, malfunctioning. These are various ways of describing systems for which the instruction is not a represented object of planning or decision-making. Or there is no capability established of obeying what is represented. Or divergence from instruction has no intentional basis to it. In order to lay out how disobedience on the part of a robot would or should work, therefore, it is important to delineate 1) what such an order or guide is, 2) what design feature the system possesses to incorporate that instruction. If there is no such identifiable order in a given situation or environment, then the robot's disobedience is only a hypothetical attribution, performing *as if* it were disobeying some order that the observer infers or makes up to put context for the robot's action. If the system has no design feature or architecture by which an order can affect its operation, then its violations or conformity toward rules out in the world are, at best, inadvertent.

The more one loosens intentions from the constitution of the action, the less disobedience can be said to differ from incidental divergence. An agent that is choosing an action for the sake of the highest reward may take the exact same action as one that is disregarding an instruction from an irrational instructor. How would one tell the difference, and when would it matter to know the difference?

One can also view the reason for disobedience as part of an implied set of counterfactuals, what needs to be the case for an instruction to be obeyed or disobeyed. If there are no conditions under which a system could both understand and competently follow an instruction, yet still not decide to do so, then one might ask if obedience and disobedience still apply. While this paper cannot explore this implication adequately, it is worth considering what degree of weighing or judging competing reasons for following an instruction is implied by obedience and disobedience both.

5 LOCAL VS. GLOBAL DISOBEDIENCE

One way to think about conflicting forms of obedience is through competing principles. Some ethical dilemmas can arise when two explicit rules cannot both be upheld, and one must account for why, and to what effect, one chooses to disobey one rather than the other. For this paper, we propose that cases of "rebel" or "disobedient" agents can be thought of as representing more or less local, and more or less global, norms and priorities. Not only do there need to be choices made as to which norm or guideline takes precedence over another, but one must ask where that precedence comes from and what enforces it. The question this embeds in matters of AI system design is whether "obedience" and "disobedience" even apply to a system designed and implemented without reference to these levels.

The better alternative to using concepts of obedience, for systems that have no internal reference to norms or ability to interact with explicit reference thereto, is to say they conform or diverge (and perhaps functions or malfunction). The intentionality at work is about what the system is "supposed" to do by the designers, without any internal deliberation about that which could be obeyed or disobeyed.

Let us consider an example of different robotic systems implemented within a hospital. The demands and challenges of COVID have only heightened questions of how robots might be useful in such a context, since they could help keep an environment sterile while performing basic tasks [10]. Imagine a delivery robot going down hallways carrying supplies from one part of the hospital to another. It has a limited natural language repertoire related to moving, stopping, and alerting others to its intended task. Let us also suppose there is a robotic system that serves as an informational kiosk, largely staying put and responding to questions from visitors.

There is no necessity of either robot being disobedient. Each robot might only respond to queries or instructions that fall within its assigned task, and lack a capability to understand or obey anything else. Perhaps by certain design elements of appearance (e.g., screen vs. no screen) visitors would not confuse the two kinds. This, again,

would be robots that were not able to obey certain instructions, meaning they do not disobey them either.

Now consider the delivery robot operating during a security emergency, when hospital policy is to have all mobile robots cease operating while it is addressed. Its stopping could be interpreted in different ways by hospital staff. Perhaps its navigation is malfunctioning, perhaps its effectors are. But if it is instructed "Take these to Room 206", it would stop from functional disobedience because of a security protocol. "This delivery cannot proceed for security" would be the more global directive that would justify its local disobedience.

Here it may be worth recalling Mirsky and Stone's "seeing-eye robot," which could disobey unsafe commands from a user who does not detect a harm that the robot perceives [17]. This would mark a difference in access to information about the world. For situations like that of delivery robot's situation, however, there might be shared knowledge about the protocol: the important point is that the robot uphold the right priority, whether it provides new information or not.

Is there a global form of disobedience for a system, where every designed standard or instruction is rejected? If there is, much less if it were sought after, then one should ask how this differs from malfunctioning, and harmful malfunctioning at that. What design reasons or purposes would such disobedience serve, if not some larger aim or principle? And if there is no question of deciding between obedience and disobedience, because there is no distinct process attached to receiving an instruction to obey, then the concept functionally falls out of the description.

Distinguishing local and global forms of disobedience allows one to compare an agent's particular actions relative to an instruction and a background norm system, allowing for modal descriptions of obedience and disobedience.. Take two norm systems A and B, where an instruction I violates A but not B. A robot trained to uphold A performs as designed if it disobeys I, whereas the robot trained to uphold B may not be obligated to follow I in order to uphold B. Robot B may disobey for local consideration (a near-simultaneous, but incompatible, instruction followed first), whereas Robot A disobeys for a global one.

6 TRANSPARENCY AND ACCOUNTABILITY FOR ASCRIPTIONS OF DISOBEDIENCE

As the introduction of self-driving cars heated up several years ago, more attention was given to how, as an autonomous system, a car should obey its owner/driver's own values and priorities. The notion of "moral proxy" is one way to describe what a locally disobedient, globally obedient system could represent, acting on behalf of an agent or community who was not directly instructed the robot in the moment [15]. The system itself, by opting to override an immediate instruction for the sake of an overarching norm, is not by virtue of that a "full ethical agent" [18]. Its upholding of that norm is more plausibly seen as a proxy for some community or societal decision about how vehicles should operate, regardless of whether a particular owner wants. Alternatively, a system that disobeys the larger societal rule for the sake of an owner's instruction would be disobedient as the owner's moral proxy. Distinguishing levels of obedience goes hand in hand with locating the moral proxy at work.

There are, of course, plenty of ways that the system could evoke overly robust ascriptions of agency, incurring blamed for its action. The manufacturers, government regulators, owners, and others could claim some malfunction was the reason for the conflict, distancing themselves from difficult decisions amid norm conflicts by branding it "going rogue". Media and entertainment harp on the them of creations turning on their creators, as many texts, from Genesis to Mary Shelley's Frankenstein, have helped summon. But these cloud the harder work of deciding what norms a robot should uphold, how design ought to achieve true accountability to them, to which norms a robot should be ultimately obedient. While liability of technical failures will always be a thorny issue to settle, designing and implementing systems without norm transparency is a serious social and technical risk. No romance of heroic rebellion or opaque algorithmic insight should obscure the importance of norms in the social fabric, as well as the ordinary demands for accountability and reason-giving that helps that fabric hang together.

The concept of disobedience implicates a role for instruction, a consideration of that instruction, and a justifiable decision to act against instruction. There are terms other than "disobedient" to describe more precisely how agents do not behave in conformity with an instruction. If an instruction is not something a system can understand or integrate into its operation – if it makes no difference to how the system choose its actions – then it is better to call it "impervious to instruction". If a system merely diverges from expected or emergent patterns (e.g., emergent coordination in a multi-agent simulation), without any explicit representation or expression of why, then it should be called "errant" (in a neutral sense of taking unexpected paths). Many impervious or errant agents could look like "disobedient rebels"; but without some feature that operationalizes an order or instruction, obedience and disobedience do not apply to them.

6.1 Obeying an instructor vs. an instruction

One response to the argument for restricting disobedience to more intentional forms is to say that, for an RL or IRL system, there is obedience and disobedience of an instructor, not so much instruction [9]. While there might not be explicitly representations, much less natural language expressions, of what is being asked of an agent, it still makes a certain amount of sense to say "Obeying or disobeying the agent just means following or not following what the instructor commands at time t". The estimation of disobedience would, then, be how to optimize a policy with an instructor whose directions might not be fully rational or accurate.

Within the confines of a dyadic relationship exploring a simple state space, this has some cogency. For more social and symbolic interactions among other agents, however, or even reasoning across time between two agents about priorities [13], the application of disobedience loses sense. Are there any rules, norms, or orderings that other agents could understand as objects of obedience and disobedience between the original two? What is the content of the disobedience being interpreted by others who would coordinate action with the artificial agent, if that agent is to be thought of as taking purposeful action rather than malfunctioning or coming up

short in its execution? If the idea is that learning reward functions from an environment may mean divergent exploring, it might make sense to call such conforming or non-conforming actions "ignoring a suggested action" and "following a suggested action". There is no independent rule being reasoned over or possibly shared with other agents, there are no explicit inferences or beliefs that can cited as reasons for its actions. In sum, the model of obeying an instructor alone is so thin a conception of sociality and rationality as to render "disobedience" without clear validity or purpose.

7 WHAT KINDS OF DISOBEDIENCE SHOULD BE SOUGHT?

Going forward it is crucial to distinguish obedience and disobedience as localized, if rudimentary, operations on the part of an interactive system from obedience and disobedience as *ascriptions* from human interactants. For systems that, as we have argued, can neither obey nor disobey by design, one could still anticipate ascriptions of obedience and disobedience by those unfamiliar with its lack of capacity. In that case one might speak of ascription mitigation, the avoidance of implementation that evokes a deluded sense that a "rebel agent" is on the loose. But that is, to repeat, quite different from actual obedience and disobedience to a concrete guideline, through a system's designed inference that a larger norm or rule is to be followed. This distinction should guide future research lest forms of "rebellion" obscure deliberate decisions as to what kind of policies, symbolic representations, and logical operators (or lack thereof) are behind the system's performance.

For the field of human-robot interaction, at least, it is worth investigating how ascriptions of disobedience detract or enhance social robotic applications. Are there more functional affordances that social robots could give, more transparent measures indicating what the system recognizes and reasons about as instruction, to deter or guide more instinctive judgments toward their actions [19]?

The point of this paper is not to sequester the term "disobedience" behind the loftiest of intentional standards. The ascription of intentionality can vary with contexts, and in some multi-agent environments it may serve as a useful heuristic to call certain systems "disobedient" so that other people do not expect instruction to be effective. But for more general discussions of accountability and design, as well as public-facing discussions of robots who elicit the label of incipient "overlords", it is important to take more care about what disobedient is really saying about the system.

The intentional invitation that "disobedience" makes to various discussions around AI and robotics is difficult to disentangle from the common sense judgments and practical reasoning that communities and societies employ. While there is no clean way to prevent the overuse of intentional terms toward artificial agents, that does not mean the invitation should go unattended and unrestrained in robotic applications. To avoid exploitative and manipulative uses of intentional language, especially a morally charged term like "disobedience", public discussions ought to reflect clarity and carefulness in technical offerings. A disobedience devoid of intentionality and without systemic transparency risks turning the intentional invitation into a provocation to fear, fantasy, and hype.

The twofold task around disobedience, then, is 1) to state what dimension of intentionality defines a system's ability to act upon a command, principle, or rule, to qualify how "disobedience" ought to be applied, 2) to map what kind of disobedience, with what form of intentionality, can most responsibly feature in a system's real-world interactions and applications.

Public discussion of robots, fueled in part by depictions of robots in film and television, often evokes the threat of rebellion or takeover from any seeming independence on the robot's part. The line "I'm sorry, Dave, I'm afraid I can't do that" from 2001: A Space Odyssey is infamous for the terror caused by a machine untethered by human command. Our discussion of disobedience, including the framework of a local vs. global disobedience, is to push against this cluster of associations. If disobedience is intentionally possible for a robot, then the difficult questions go to its design and how its intentions conform to the priorities of those affected by it. If it is not truly disobedient, then the responsible question is to what degree it can take instruction – its supposed obedience and disobedience may be illusory, and its disconnection from human instruction even more recklessly severe.

8 CONCLUSION

The idea of disobedience from artificial agents carries two important points for design. First, disobedience is a term with intentional implications and connotations, and ignoring these can easily misrepresent what the system is doing and how it is to be held accountable. A system that disobeys is one that is equipped to obey, capable of obeying, but has an accessible reason to take an alternative course of action. Second, for a system to disobey responsibly, the reasons to obey or disobey must be specified and ordered in an accountable fashion. The priorities, norms, and commitments that take precedence over others, however rudimentary they are in context, should define the need for and function of disobedience. Impervious or errant systems may learn to navigate the world, but their lack of conformity is not an intentional disobedience; consequently, the design of a properly and usefully *disobedient* robot must integrate and offer accurate access to reasons. Exaggerating and overpromising intentionality does not just fuel "rogue robots" hype on the way to public declamations of panic, it fundamentally attacks an interactive norm of transparency and accountability. If robots are to disobey for reasons that matter, reasons must matter in their decisions.

REFERENCES

[1] 2022. The Runaway Robot: How One Smart Vacuum Cleaner made a break for freedom. https://www.theguardian.com/lifeandstyle/2022/jan/24/the-runaway-robot-how-one-smart-vacuum-cleaner-made-a-break-for-freedom

[2] Ronald Arkin. 2009. *Governing lethal behavior in autonomous robots.* Chapman and Hall/CRC.

[3] Isaac Asimov. 1942. Runaround. *Astounding science fiction* 29, 1 (1942), 94–103.

[4] Nick Bostrom. 2014. *Superintelligence. Paths, Dangers, Strategies.*

[5] Gordon Briggs and Matthias Scheutz. 2015. "Sorry, I can't do that": Developing Mechanisms to Appropriately Reject Directives in Human-Robot Interactions. In *AAAI Fall Symposium Series: Artificial Intelligence and Human-Robot Interaction.*

[6] Gordon Briggs and Matthias Scheutz. 2017. The case for robot disobedience. *Scientific American* 316, 1 (2017), 44–47.

[7] Alexandra Coman and David W Aha. 2018. AI rebel agents. *AI Magazine* 39, 3 (2018), 16–26.

[8] John H Flavell. 1988. The development of children's knowledge about the mind: From cognitive connections to mental representations. (1988).

[9] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. 2016. Cooperative inverse reinforcement learning. *Advances in neural information processing systems* 29 (2016).

[10] Jane Holland, Liz Kingston, Conor McCarthy, Eddie Armstrong, Peter O'Dwyer, Fionn Merz, and Mark McConnell. 2021. Service robots in the healthcare sector. *Robotics* 10, 1 (2021), 47.

[11] Charles W Kalish and Rebecca Cornelius. 2007. What is to be done? Children's ascriptions of conventional obligations. *Child Development* 78, 3 (2007), 859–878.

[12] Daniel Kasenberg and Matthias Scheutz. 2017. Interpretable apprenticeship learning with temporal logic specifications. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*. IEEE, 4914–4921.

[13] Daniel Kasenberg and Matthias Scheutz. 2018. Norm conflict resolution in stochastic domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[14] Séverin Lemaignan, Julia Fink, Francesco Mondada, and Pierre Dillenbourg. 2015. You're doing it wrong! studying unexpected behaviors in child-robot interaction. In *International conference on social robotics*. Springer, 390–400.

[15] Jason Millar. 2015. Technology as moral proxy: Autonomy and paternalism by design. *IEEE technology and Society Magazine* 34, 2 (2015), 47–55.

[16] Smitha Milli, Dylan Hadfield-Menell, Anca Dragan, and Stuart Russell. 2017. Should robots be obedient?. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 4754–4760.

[17] Reuth Mirsky and Peter Stone. 2021. The Seeing-Eye Robot Grand Challenge: Rethinking Automated Care. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*. 28–33.

[18] James Moor. 2009. Four kinds of ethical robots. *Philosophy Now* 72 (2009), 12–14.

[19] Robert A Paauwe, Johan F Hoorn, Elly A Konijn, and David V Keyson. 2015. Designing robot embodiments for social interaction: affordances topple realism and aesthetics. *International Journal of Social Robotics* 7, 5 (2015), 697–708.

[20] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M Howard, and Alan R Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*. IEEE, 101–108.

[21] Stuart Russell, Daniel Dewey, and Max Tegmark. 2015. Research priorities for robust and beneficial artificial intelligence. *Ai Magazine* 36, 4 (2015), 105–114.

[22] Matthew Sparkes. 2022. Tesla recalls 50,000 cars that disobey stop signs in self-driving mode. *New Scientist* (2022). https://www.newscientist.com/article/2307147-tesla-recalls-50000-cars-that-disobey-stop-signs-in-self-driving-mode

[23] Susumu Tachi and Kiyoshi Komoriya. 1984. Guide dog robot. *Autonomous mobile robots: Control, planning, and architecture* (1984), 360–367.