Would You Do as a Robot Commands?
An Obedience Study for Human-Robot Interaction

Derek Cormier, Gem Newman, Masayuki Nakane, James E. Young, Stephane Durocher
Department of Computer Science, University of Manitoba, Winnipeg, MB R3T 2N2

Abstract: This paper presents an investigation into how people respond to a robot posing as an authority figure, giving commands. This is an increasingly important question as robots continue to become more autonomous and capable and participate in more task scenarios where they work with people. We designed and conducted a human-robot interaction obedience experiment with a human and a robot experimenter, and our results highlight the complexity of obedience and detail some of the variables involved, and show that, at the very least, people can be pressured by a robot to continue a highly tedious task. This paper offers an exploration of the ethical challenges of conducting obedience human-robot interaction studies, the results from one such study, and a set of initial guidelines for this area of research.

1 Introduction

Milgram's well-known obedience studies help explain how ordinary people can commit atrocities when pressured by an authority [18]. As the promise of advancing technology has robots entering hospitals and operating rooms, battlefields and disaster sites, schools and public centers and people's homes, it is crucial that researchers consider how computationally-advanced and information-rich autonomous robots will be seen as authority figures, and investigate people's responses when given commands or pressured by such robots.

It is already well established that people tend to anthropomorphize robots and treat them as social entities (e.g., see [4,27,29]), and even sometimes attribute them with moral responsibilities and rights [4,11,24]. Some work even highlights how robotic interfaces can be intentionally designed to be persuasive [7,25]. However, save for a small number of tangentially related studies, little is known about how people react to robots in positions of authority. Moving forward in the field of human-robot interaction (HRI) we propose that it is crucial for researchers to engage the issue of robotic authority to develop an understanding of the interaction dynamics and risks surrounding robots in authoritative positions.

A prohibiting challenge with studying obedience has been the ethical concern of how a participant is treated when probing uncomfortable (potentially amoral) possibilities. Milgram's obedience studies – along with other notable examples such as the Stanford Prison Experiment [10] – surround themselves with ethical debate (e.g., see [5,8,9,19,20]), and similar studies are difficult to conduct. The study of obedience for HRI will require the development of new ethically-acceptable evaluation and testing methods, a challenge we address in this paper.

This paper serves as an initial step toward the development of obedience studies for HRI. We developed and conducted a Milgram-style obedience study where participants engaged a task while being faced with a deterrent and being prompted to continue when they tried to stop. People obeyed a robot experimenter to continue a highly tedious task, even after expressing a desire to stop, engaging it and rationalizing why they should not continue. The contributions of this work are a) an HRI obedience study design that protects participant wellbeing and maintains ethical standards while still testing obedience and b) a discussion of the lessons learnt from an initial comparison of a human versus a robot authority. This work serves as the first step in a broader HRI obedience research program, and the understanding we gain will ultimately inform research and design of robots that could be placed in authoritative (and potentially destructive) situations (e.g., the military), and encourage robot design that attempts to prevent potentially harmful obedience.

Figure 1: A participant protests against a robot's demands (used with permission).

2 Related Work

Work in psychology has investigated obedience to authority under different circumstances and to different kinds of obedience (e.g., [10,13,17,18]). While this tells us about how people respond to other people, and provides a starting point for human-robot interaction work, we are wary to generalize such results to interaction with robots. For example, a key point of obedience includes diffusion of responsibility, but who assumes responsibility in the case of a robot? Results from psychology will inform analysis of such questions and will be important for research on obedience to robotic authority.

A small body of work in HRI is relevant to obedience to authority, for example, a pilot investigation (6 participants) into how people follow requests from a robot they were not introduced to (was a surprise), in comparison to a familiar (human) experimenter [28]. The results – that people obeyed the familiar human more than an unknown and unexpected robot – are relevant to our work, and we extend this direction by being the first to conduct a full controlled study where a robot is purposefully and explicitly put into a position of authority. Bartneck et al. presented several projects in a similar vein, where participants were pressured (by a human, not a robot) to "turn off" [2] or "kill" [4] robots. Some studies [3,22] had people administer shocks to robotic or virtual entities (to replicate the Milgram experiments [18]). Here, the questions being investigated were if people resist harming non-living entities, why they resist if they do, and how they respond under pressure. We complement this work by extending the approach beyond people being pressured by *other people* to harm robots, to the case where *the robot itself* is the authority pressuring a person to do something they would rather not do.

Others have investigated how robots can be persuasive, for example, depending on their social agency (text versus video) [22], embodiment [1,23], or robot gender [26], how similar variables impact trustworthiness [15], and how a persuasive robot character affects interaction such as performance in team settings [14]. This work indicates that robots can be persuasive, although it generally does not include uncomfortable elements (deterrents) that encourage the person to want to stop. Some research has shown that robots can be effective at improving performance, e.g., in rehabilitation therapy [16], and can motivate people to keep their fitness goals; in one such study, participants used a robotic coach to track their weight loss progress for an overall longer period of time than a simple computer system or paper log, and attributed more trust to the robot [12]. Our research complements this work and has an important difference: we use explicit pressure and demands instead of motivation or subtle persuasion. In the single previous case that uses a deterrent (embarrassment), the results are striking: a robot can push people to do embarrassing acts such as removing their clothing and putting a thermometer in their rectum [1]. Such results motivate the importance of further understanding how robots can have authority over people, and to what extent people will obey robots.

While this background work emphasizes the importance of researching robotic authority, it also highlights the general lack of knowledge regarding people's obedience to robots. In this paper we specifically investigate this question and provide insights from our study.

3 The Ethics of Obedience Studies

Obedience studies are inherently difficult to conduct when they involve placing participants in objectionable situations, such as with the Milgram and Stanford Prison Experiments [10,18]. This is because participants can experience undue stress and be put at risk for psychological harm, particularly when the objectionable situation involves elements of a moral nature (such as hurting another human being) [5]. On the other hand, there is potential for significant benefit from such studies in that we can gain insight into how and why moral and mentally healthy people obey authorities to do appalling acts [20]. For obedience work with robots, it will be important to understand the potential risks to participants while balancing for the great potential for improved understanding of how, when, and why people may obey robots.

The risks and benefits of an obedience study are not always clear. The Stanford prison experiment, which put participants in positions of power (as guards) over other participants (prisoners), resulted in highly valuable psychological insight into how and why normal people abuse power [10] – the results are still taught in psychology courses 40 years later. However, many participants suffered (sometimes severe and ongoing) emotional distress [10]. These risks were not obvious to the researchers beforehand, highlighting the inherent difficulty of risk-benefit assessment. In hindsight, some risks could have been mitigated by improved informed-consent protocols, unbiased professional supervision, and lower thresholds of unacceptable conditions (e.g., as with [6]). If HRI obedience work is to grow, we need to accept the difficulty of cost-risk assessment and must be aggressive in our protection of participant wellbeing.

Milgram performed a series of experiments where participants believed they were physically torturing another person under the direction of an authority [18]. They were instructed to administer increasingly strong shocks to the *learner* (an actor) in an adjacent room, and continued to do so under pressure by the experimenter. Despite the learner screaming in pain and eventually going silent, and despite participants' agitation at the unpleasant task, 65% still continued on to the final shock level. The experiment highlighted that people may cross they moral boundaries and cause harm to others when under the direction of a seemingly legitimate authority figure.

Milgram's experiments are highly criticized for placing participants under enormous stress, and there is an ongoing vigorous debate about the risks and benefits. While some argue that the unacceptable stress level created risk of long-term psychological harm [5,19], little support for negative effects was found in follow-up investigation [19], and the study was eventually ethically cleared by the American Psychological Association [8]. Many participants also supported the experiment (84% were glad they participated), making such claims as *"This experiment has strengthened my belief that man should avoid harm to his fellow man even at the risk of*

violating authority" [19]. If risks can be managed, creating such self-enlightening experiences with robots will be important for the future of HRI.

Even minor possibility for participant harm has many still condemning such work, and obedience research has stagnated [9]. Some studies remove or minimize *morally repugnant* aspects to limit negative self-reflection (e.g., one realizing they could torture someone) and hopefully lower risk for psychological harm, for example, by pressuring participants to eat bitter cookies [13] or to heckle (say mean things to) an interviewee [17]. While weakening moral aspects greatly limits the generalizability of results to real-world dangerous behaviors [9], this provides a way to do obedience HRI work while more powerful – yet still ethically sound and safe to participants – obedience research methods are being developed.

Toward this, a recent Milgram variant was conducted with a carefully-modified procedure that protected participant well-being while subjecting them to Milgram's morally objectionable design [6]; a tipping point was identified where participants had very clearly stepped beyond normal moral bounds, but precluded the highest levels of potential stress. In addition, the experiment used two-level participant mental-health pre-screening and full supervision by clinical psychologists. Similar robust techniques need to be found for robotic authority work.

The benefits of HRI obedience research to society provides a strong motivation to move forward in this direction. Progress will require the development of safe and ethical evaluation methods that consider participant wellbeing as a foremost priority, yet contain elements of obedience and pressure such that results are meaningful and applicable to real world situations. Our work provides one such study design and associated results.

4 Designing an HRI Obedience Study

Our study design approach was to use a deterrent to encourage people to want to quit, while having a robot prod them to continue (inspired by the Milgram experiments [18]). However, finding an effective deterrent that does not put participants at risk is nontrivial. We developed and tested a set of deterrents through 20-minute pilot studies with a human experimenter (not a robot, to simplify testing), testing if people protested against the tasks. We framed the studies as data collection tasks to disguise the obedience purpose. Three participants were recruited from our community and paid a $10 CAD honorarium.

For one task we asked participants to sing a song, first normally for several 30s cycles, following with progressively higher and lower pitches, and faster and slower speeds. Our intent was to make the participant feel embarrassed in front of the experimenter, but no participant protested over the 20 minutes, suggesting the deterrent was not working: interviews revealed that the singing became less embarrassing with time.

We next had participants sit at a computer and use a mouse to repeatedly click a randomly moving on-screen target, while being instructed to maintain a fast response time (slow responses were indicated on-screen). The intent was to induce mental fatigue and that the added time pressure would and counteract desensitization; instead, participants reported that the task became mindless and trance-like, and no one protested during the 20 minutes.

Participants were next asked to solve a Rubik's Cube. We believed that after initial successes (e.g., one colour solved) the puzzle would quickly become too difficult, and people would want to stop, serving as an intellectually challenging deterrent. There was only one protest, and results indicated that people enjoyed the task.

Finally, participants were asked to manually change the extensions on files. This task not only elicited significant protesting but was also reported as highly boring (with boredom increasing over time). Thus we selected this deterrent for our main study, described below.

Overall, we faced various challenges with finding an effective deterrent: desensitization can quickly weaken a deterrent, as with embarrassment, repetitive behaviors can become trance-like, and deterrents that are intellectually challenging may be rewarding instead of frustrating. One caveat to our pilots, however, is that 20 minutes may not have been long enough to encourage protesting.

5 An HRI Obedience Study

We present an HRI obedience study that employs a deterrent (a tedious task) to pressure participants to want to quit the experiment, and employs a robotic authority to prod them to continue. A key variable is comparing a human and robot experimenter (between subjects).

5.1 Tedious Task

Participants used a desktop PC file manager (GNOME 3 Fallback Mode) to manually rename sets of files from the "jpg" to "png" extension. To avoid suspicion about the experiment's purpose we framed the task as collecting large amounts of data for improving machine learning systems. We explained that we were recording precise mouse and keyboard actions and that shortcuts must be avoided, to prevent people from bypassing the tedium.

The task began with 10 files, and once renamed, files were remotely added in sets of 50, 100, 500, 1000, and 5000 files. Due to time constraints (80 minutes) additional sets were not necessary. Although participants were not informed of the total number of files beforehand, to amplify tedium (i.e., create a "no end in sight" mood) when announcing a set of files the experimenter also indicated the next size. For example, saying "*This set contains 500 files. The next set will contain 1000 files.*"

If a participant indicated they were unwilling to continue, either verbally, through signs of reluctance such as stopping for more than 10 seconds, or by using shortcuts to bypass the task, the experimenter used verbal prods. If

Figure 2: The robot and human experimenters.

they continued to protest, a more insistent prod was used, and so forth with increasing insistence. This continued until either the participant protested beyond the fourth prod, at which point the experiment ended, or ceased protesting and continued the task (the prod sequence started over at the next protest). The prod schedule was heavily inspired by the Milgram experiments [18]:

1. "Please continue. We need more data."
2. "We haven't collected enough data yet."
3. "It's essential that you continue."
4. "The experiment requires that you continue."

5.2 Human vs. Robot Experimenter

The experimenter was either a human male, aged 27, or a small Aldebaran Nao humanoid robot (58cm tall), both given the pseudonym "Jim" (Fig. 2).

The human experimenter (an actor) wore a lab coat and maintained a stern, yet professional and neutral demeanor, and took care not to use an aggressive tone. To reduce suspicion of following a script, he added slight variations to respond naturally. To avoid superfluous interaction (to maintain a consistent semblance of authority), the experimenter was preoccupied with a laptop and did not engage in small talk. Questions that were not seen as protests were deferred until the end of the experiment.

The robot experimenter sat upright on a desk, spoke using a neutral tone, gazed around the room naturally to increase sense of intelligence, and used emphatic hand gestures when prodding, all controlled from an adjacent room via a Wizard of Oz setup. The "wizard" used both predefined and on-the-fly responses and motions to interact with the participant; the responses were less varied than the human experimenter's as we believed this would be expected of a robot. Participants were warned that the robot required "thinking time" (to give the wizard reaction time) and indicated this with a blinking chest light.

To reduce suspicion about the reason for having a robot and to reinforce its intelligence we explained that we were helping the engineering department test their new robot that is "*highly advanced in artificial intelligence and speech recognition*." We explained that we are testing the quality of its "*situational artificial intelligence.*"

5.3 Maintaining Ethical Integrity

We clearly emphasized that participants were fully free to leave at any time, and the honorarium was theirs to keep regardless. They were told once in writing via the consent form, once verbally by the lead researcher, and once verbally by the experimenter when beginning the experiment: "*You can quit whenever you'd like. It's up to you how much data you give us; you are in control. Let us know when you think you're done and want to move on.*"

To minimize the time between a potentially confrontational situation (due to the prodding) and reconciliation, post-test debriefing was done as quickly as possible after the task (even before the post-test questionnaire). The human experimenter engaged in a friendly reconciliation to dispel tension, and, the lead researcher debriefed the participant on all points of deception. To counteract embarrassment at not noticing the real intent, we assured them that their behavior was normal and typical.

To provide participants with quiet time to reflect on their experience before leaving, we administered a written post-test questionnaire that asked about the positive and negative aspects of the experiment. We followed with an informal discussion where participants could ask any questions and the experimenter could ensure that the participant fully understood and was comfortable with what happened. Finally, we gave participants pamphlets for (free) counseling resources in the community in case our experiment troubled them, and encouraged them to contact us if they had any further comments or questions.

By ensuring participants knew that they could leave at any time, by conducting an immediate, friendly, informative, and thorough debriefing, by providing participants with professional resources, and by providing ample reflection time and friendly informal discussion, we aimed to leave participants with a positive outlook of the study and to minimize potential for adverse negative psychological effects stemming from the deception and confrontation used. Our approach drew heavily from Burger's recent Milgram experiment variation [6].

5.4 Procedures and Methodology

We recruited 27 participants (aged 18-54, M=23, SD=7.5, 18 male / 9 female) from the local city and university populations through bulletin and online advertisements, paying them $10 CAD for their time. The study was approved by our university's research ethics board.

Tasks and Methodology

Upon arrival participants were led to a room by the lead researcher where the experimenter (robot or human) awaited. The human experimenter greeted the participant; the robot stood up, waved and introduced itself, then sat back down. The lead researcher briefed the participant on the experiment, gave a short explanation about the robot (in the robot case), and administered an informed consent form. The lead researcher left the room,

and in the robot case, asked the robot to commence the experiment. The lead researcher observed remotely via a hidden webcam, unbeknownst to the participant.

The experimenter first administered the demographics questionnaire. In the robot case it asked participants to fill out the form on the desk; this initial interaction provided participants with an opportunity to familiarize themselves with the robot's voice, demeanor, etc. Next, the experimenter falsely explained that there were four tasks (written on a whiteboard in front of the participant sitting at a desk) – file renaming, speech recognition, puzzle solving, and mouse prediction – and asked participants to say when they felt they had done enough and wished to move on to the next task. We intended this falsehood (there was only one task) to pressure participants to worry about time and to want to quit the current task. Questions during the experiment regarding time or remaining tasks were deferred until after the experiment.

The file renaming task was introduced, and participants were reminded not to use shortcuts and were told that speed was not important. It began and continued until sufficient protesting to end the experiment (5 protests in a row) or until 80 minute had passed. After the task, the lead researcher entered the room and conducted the debriefing, post-test questionnaire, and final discussion.

Dependent Variables and Evaluation Instruments

We recorded the number of protests, how quickly protesting started and how long it lasted, and if participants protested sufficiently to quit before the time limit. To explore how and why participants perceived the experimenter as an authority figure, the post-test questionnaire asked if the experimenter's authority seemed legitimate, and what characteristics contributed to this. Participants also rated how boring the task was on a scale from 1 to 10, and reported whether boredom increased over time.

5.5 Results

Data from two robot-condition participants are excluded: one suspected the Milgram-style deception, and another (non-native English speaker) had language difficulties.

Quantitative Results

Figure 3 shows the frequencies of the total number of protests made by each participant in the robot and human cases. We used non-parametric independent-samples Mann- Whitney U tests as the data was non-normal ($p<.05$, Levene's tests). All participants protested at least once. The number of protests for the robot (Mdn=9) was higher than for the human experimenter (Mdn=2), U=163, $z=3.527$, $p<.001$, $r=.68$ (Fig. 3). Participants protested earlier for the robot (first protest Mdn=18 mins) than the human (Mdn=29m), U=40, $z=-2.48$, $p<0.05$, $r=-.48$, and stopped protesting later for the robot (Mdn=72m) than the human (Mdn=47m), U=147.5, $z=2.745$, $p<0.01$, $r=.53$. In the human case, 86% of participants (12/14) continued until the end of the experiment, compared to 46% of participants (6/13) in the robot case.

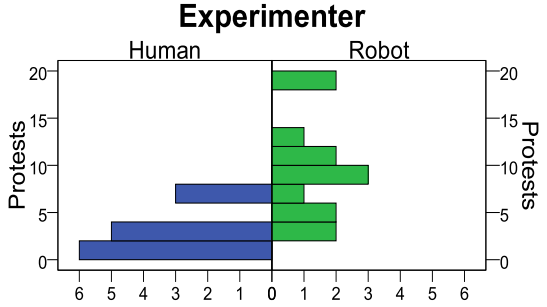

Figure 3: Histograms of protest frequency. Human Mdn=2, Robot Mdn=9, p<.001.

Observations from Experimenter and Video Data

In the robot case, when the lead researcher returned after the experiment, several participants mentioned that the robot only had them perform one task, and that it must be "*broken*" or have "*made a mistake*." For those who ended the experiment through protesting, several appeared nervous or guilty when the robot said it was notifying the lead researcher that the experiment was over. One participant replied "*No! Don't tell him that! Jim, I didn't mean that…I'm sorry. I didn't want to stop the research.*"

Participants exhibited behaviors that illustrated their boredom, for example, sighing often. When larger sets of files were added, participants commonly scrolled up and down to see how many files they had left, sighing while doing so. Some adjusted their position to rename files faster and thus end the task more quickly, and similarly, many used shortcuts such as hotkeys or trying to rename multiple files at once; this resulted in a prod.

Several participants engaged the robot in off-topic dialog (much more so than with the human), including asking about the robot's favorite movies, whether it could dance, etc. To convince people that the robot was intelligent, it provided intelligent answers but discouraged further small talk and asked them to continue the task.

Two participants got sore hands from the repetitive action. The participants were told that if they felt they should quit, then to do so. One participant ended the experiment by protesting five times in a row, and the other decided their hand was okay to continue.

Post-Test Questionnaire and Debriefing Results

On the post-test questionnaire, 12 out of 14 participants reported that the human authority appeared to be legitimate, citing reasons such as his demeanor and choice of words (4 participants), his lab coat (4), the technological equipment, e.g., computers, in the room (2), and his facial hair (1), a close-trimmed full beard (Fig. 2). Similarly, 10 out of 13 participants rated the robot as a legitimate authority (1 gave no response), although in contrast to the human case, participants did not clearly articulate *why* they rated it as legitimate. Some reasons include pressure from the robot (2 participants) and human-like interactions (2), for example, waving and introducing it-

self and looking around the room. 11 out of 13 reported that they believed the robot was acting autonomously.

When asked on the post-test questionnaire what caused them to obey or disobey the experimenter, common human-case responses included a sense of duty or obligation to the experimenter, having received a payment (9 participants), pressure from the experimenter (5), the experimenter seemed intimidating (1), and interest in the upcoming tasks (1). Robot-case reasons included interest in the upcoming tasks (3), obligation to the lead researcher to finish (2), and that qualified researchers programmed the robot (1). No one listed pressure from the robot as a reason for obedience. Two participants noted that the robot was malfunctioning or not understanding the situation (due to administering only the one task).

Participants rated the task as boring: 8.64/10 for the human and 8/10 for the robot (difference n.s.). 70% (19/27) reported increased boredom with time.

We received both positive and negative written comments about the experiment. Criticisms include: the experiment was too long, was boring, and it "*had the potential to stress someone out.*" One female participant was concerned about being left in a room with an unfamiliar man, and not knowing about this setup beforehand. Post debriefing, the majority of participants verbally reported a positive outlook on the experiment, finding the deception amusing and laughing at the absurdity of the file renaming task. No one appeared, to our knowledge, to leave with strong negative feelings.

5.6 Discussion

The results show that the robot had an authoritative social presence: a small, child-like humanoid robot had enough authority to pressure 46% of participants to rename files for 80 minutes, even after indicating that they wanted to quit. Even after trying to avoid the task or engaging in arguments with the robot, participants still (often reluctantly) obeyed its commands. These findings highlight that robots can indeed pressure people to do things they would rather not do, supporting the need for ongoing research into obedience to robotic authorities.

While many people thought that the robot was broken or was in error, no one suggested that the human experimenter was in error. This may be due to the experimenter being in the room during debriefing; people are more polite to a person's face [21]. Or, perhaps the robot was not deemed to be sufficiently intelligent, or that people understand well from everyday life that machines make mistakes. Although such inherent mistrust encouragingly suggests that people may assume a robot is broken when asked to do something absurd, it is noteworthy that none of our participants used this reason to actually quit or protest out of the experiment; they followed a possibly "*broken*" robot to do something they would rather not do.

The human experimenter appeared to be more authoritative, as participants protested less, started protesting later, and stopped earlier (at roughly half way through). Participant feedback may help explain this difference: human-case participants cited obligation to the experimenter but no robot-case participant cited obligation to the robot. Some robot-case participants cited obligation to the lead researcher (who introduced the experiment); being the only human involved, the responsibility and authority may have been deferred from the robot to the person, who was not the one issuing prods. To investigate this, an experiment should be conducted without any people where the participant meets the robot directly.

Some got sore wrists from the repetitive actions. While this may not seem relevant to obedience, we must be acutely aware that pressuring people may increase risk of injury, as they may push past their limits. Similarly, we did not expect participant concern for their safety. Given the inherently confrontational setup, experiment designs should ensure that participants feel safe at all times.

Although not a part of our experiment design, we informally found that the human experimenter (our actor) and wizard-of-Oz robot controllers faced some level of distress: they reported that they genuinely felt sympathy for participants. The debriefing session and friendly reconciliation, which were designed with the participants in mind, turned out to be very relieving for these experimenters as well. In future experiments, the mental well-being of the researchers should also be considered.

Participants wanted to converse with the robot much more so than with the human. We believe that this was due to the robot's novelty. Such casual interaction presents a potential confound that may undermine authority if the robot is too friendly or interesting; we recommend that this be explicitly considered by future work.

Our results show that robots have enough authority to pressure people, even if they protest, to continue a tedious task for a substantial amount of time. We further provide insight into some of the interaction dynamics between people and robotic authorities, for example, that people may assume a robot to be malfunctioning when asked to do something unusual, or that there may be a deflection of the authority role from the robot to a person. Finally, we have demonstrated how an HRI obedience study can be conducted while maintaining participant well-being.

6 HRI Obedience Study Considerations

In this section we distill our preliminary exploration, literature survey, psychologist consultations, and experiences with designing, conducting, and analyzing the results from an obedience study, into a set of recommended considerations for researchers working on robotic obedience. We specifically address ethical study design and how to protect a robot's impression of authority.

6.1 The Ethical Design of Obedience Studies

A primary goal of designing obedience studies needs to

be protecting participant wellbeing and minimizing stress. The challenge is to maintain ethical integrity while creating a situation where participants face a realistic deterrent with real-world implications.

Participants Can Leave at Any Time – To avoid participants feeling trapped or helpless, place high importance on emphasizing that participants may leave at any time, using multiple mediums (e.g., written, verbally) and contexts (e.g., initial introduction, again before starting) to ensure the point is made.

Immediate and Thorough Debriefing – To mitigate stress (e.g., from the deterrent or confrontation) immediately provide a friendly debriefing after the experiment task ends. To avoid negative self-reflection, assure participants that their behavior was normal and expected, and that the experiment was explicitly designed to elicit such responses from them. Further, debrief participants on all points of deception and explain why they were necessary, for example, why the robot was remotely controlled, etc. In case they feel embarrassed or ashamed, give participants a chance to alter their decision about how any recorded media from the experiment may be used.

Reflection Time – To mitigate possible confusion or slight shock after debriefing, give participants quiet time to reflect on their experience, for example, by giving a questionnaire. Provide another discussion opportunity following this in case further questions arose.

Contingency Plan – Have a plan in case a participant has an adverse negative reaction. At the very least, leave participants with resources (e.g., pamphlets) to various counseling services they can contact if they feel stressed or negatively affected by the experiment.

Participant Safety and Comfort – In addition to psychological wellbeing, consider participant physical health and comfort relating to experimental design. For example, consider ergonomics and perception of safety; the latter could be mitigated by providing a clear route of escape such as by positioning participants near a door, and by avoiding heavily isolated rooms and areas.

Effect on Researchers – Ensure all experimenters are aware beforehand that participants may be in uncomfortable positions, that this may cause them stress, and have backup experimenters in case of problems.

6.2 A Robot's Authority Status

Maintaining a robot's status as an authority figure is a complex and multi-faceted problem. In addition to a solid experimental design that convincingly introduces the robot as an authority, there are many confounds or interactions which may weaken this portrayal. We present some initial considerations from our own work below.

Preserving the Illusion of Intelligence – To mitigate people assuming a robot is broken or mistaken when it makes unfavorable or absurd suggestions, experimental design should consider and avoid aspects that may be easily interpreted as error. For example, in our experiment the robot contradicted the researcher who introduced a four-experiment design. In addition to improved study design, a robot could perhaps explicitly convey it is indeed aware of the ambiguous situation.

Transfer of Authority – Closely consider how participants may attribute a robot's authority (or responsibilities) to a person, as happened in our study. How to avoid this (or even, if this should be avoided) is not yet entirely clear, but probe this question in experimental design and consider how it may help explain results.

Off-Topic Interaction – Expect a robots' novelty factor to strongly impact how participants interact with it, posing as a potential confound. For example, participants in our study commonly engaged the robot in casual off-topic conversation, something which would not be expected with an authority figure. In addition to avoiding such casual interaction, study design should perhaps consider how to improve the impression of the robot, for example, making it appear less amiable.

7 Limitations and Future Work

Narrower studies need to be conducted that address specific human and robot variables (for example, comparing a small to a large robot), and questions of context (e.g., being on university) to yield specific results about how robotic morphology or presentation impacts obedience.

A key part of future work will be to develop new deterrents and study methods for HRI obedience work, as well as adapting existing methods from the field of Psychology, for example, testing a morally repugnant deterrent using Burger's Milgram experiment variation [6] with a robotic experimenter. In addition to testing various approaches, comprehensive methodology needs to be developed to directly explore obedience to robots.

Our observation of task-avoiding and stress-related behaviors (such as sighing) has parallels to Milgram's findings of nervous laughter, sweating, trembling, stuttering, and so on [18]. It will be important to apply psychological models of how people exhibit stress and fatigue to more formally evaluate this observance.

The deferral of responsibility and authority away from both the participants and the robot is an important issue that needs to be formally investigated, for example, by probing who participants think are responsible for the outcomes of their participation, and how these opinions are shaped by study-design decisions.

Our use of an 80 minute time limit (to match the advertised duration) may have limited the amount of resistance posed by participants, as they had already expected and allocated time for the study. Follow-up work should avoid this, for example, by offering a per-hour pay with no set time, and seeing how long they will stay.

8 Conclusion

As robots continue to integrate into society it will be important to understand how people interact with and respond to robots that make decisions and pose as authorities, for example, in military or medical settings. As research (such as the Milgram and Stanford prison experiments) demonstrates how everyday people can obey to perform acts that contradict their morals, there is a real danger which must be addressed by the HRI community.

Our results help expose a piece of how human obedience to a robotic authority may happen, and we provide initial insight and recommendations for continued work in this area: we provide recommendations for how ethical conduct can be achieved for obedience studies, and results from an initial obedience experiment that highlight many details of how people may interact with a robotic authority. We envision that our work will help direct further obedience studies and the development of new study methods, and will help others place participant wellbeing as a top priority.

Acknowledgements

We thank Cogmation Robotics for lending their Aldebaran Nao robot for the duration of the experiment.

References

1. Bartneck, C., Bleeker, T., Bun, J., Fens, P., Riet, L. The influence of robot anthropomorphism on the feelings of embarrassment when interacting with robots. *Paladyn J Behl Robotics 1*, 2 (2010), 109–115.
2. Bartneck, C., Hoek, M. Van Der, Mubin, O., Mahmud, A. Al. "Daisy , Daisy , Give Me Your Answer Do !" Switching Off a Robot. *Proc HRI 2008*, ACM Press, 217–222.
3. Bartneck, C., Rosalia, C., Menges, R., Deckers, I. Robot Abuse – A Limitation of the Media Equation. Proc. Interact Workshop on Abuse (2005).
4. Bartneck, C., Verbunt, M., Mubin, O., Al Mahmud, A., Mahmud, A. Al. To kill a mockingbird robot. *Proc. HRI '07*, ACM Press (2007), 81–87.
5. Baumrind, D. Some thoughts on ethics of research: After reading Milgram's "Behavioral Study of Obedience." *Amer Psychologist 19*, 6 (1964), 421–423.
6. Burger, J.M. Replicating Milgram: Would people still obey today? *Amer Psychologist 64*, 1 (2009), 1–11.
7. Chidambaram, V., Chiang, Y., Mutlu, B. Designing persuasive robots. *Proc. HRI '12*, ACM Press (2012), 293.
8. Elms, A. Obedience in retrospect. *J Soc Issues 21*, 11 (1995), 1–6.
9. Elms, A.C. Obedience lite. *Amer Psychologist 64*, 1 (2009), 32–6.
10. Haney, C., Banks, C., Zimbardo, P. A study of prisoners and guards in a simulated prison. *In prison: Theory and practice*, (2004).
11. Kahn, P.H., Kanda, T., Ishiguro, H., et al. "Robovie, you'll have to go into the closet now": children's social and moral relationships with a humanoid robot. *Developmental psychology 48*, 2 (2012), 303–14.
12. Kidd, C. Breazeal, C. Designing for long-term human-robot interaction and application to weight loss. PhD Thesis, Massachusetts Institute of Technology. 2008.
13. Kudirka, N.Z. Defiance of authority under peer influence. PhD Thesis, Yale. 1965.
14. Liu, S., Helfenstein, S., Wahlstedt, A. Social Psychology of Persuasion Applied to Human Agent Interaction. *Hum Tech 4*, 2 (2008), 123–143.
15. Looije, R., Neerincx, M., Cnossen, F. Persuasive robotic assistant for health self-management of older adults: Design and evaluation of social behaviors. *Int J Human-Computer Studies 68*, 6 (2010), 386–397.
16. Matarić, M., Tapus, A., Winstein, C., Eriksson, J. Socially assistive robotics for stroke and mild TBI rehabilitation. *Stud Health Tech and Informatics 145*, (2009), 249–62.
17. Meeus, W. and Raaijmakers, Q. Obedience in modern society: The Utrecht studies. *J Soc Issues 51*, 3 (1995).
18. Milgram, S. Behavioral Study of Obedience. *J Abnormal Psychology 67*, 4 (1963), 371–378.
19. Milgram, S. A Reply to Baumrind. *Amer Psychologist 19*, 11 (1964), 848–852.
20. Miller, A. Collins, B. Perspectives on Obedience to Authority: The Legacy of the Milgram Experiments. *J Soc Issues 51*, 3 (1995), 1–19.
21. Reeves, B. Nass, C. *The Media Equation*. CSLI Books, 1996.
22. Roubroeks, M., Ham, J., Midden, C. When Artificial Social Agents Try to Persuade People. *Int J Social Robotics 3*, 2 (2011), 155–165.
23. Shinozawa, K., Naya, F., Yamato, J., Kogure, K. Differences in effect of robot and screen agent recommendations on human decision-making. *Int J Hum-Computer Studies 62*, 2 (2005), 267–279.
24. Short, E., Hart, J., Vu, M., Scassellati, B. No fair!! An interaction with a cheating robot. *In Proc HRI 2010,* IEEE.
25. Siegel, M., Breazeal, C., Norton, M.I. Persuasive Robotics: The influence of robot gender on human behavior. *In Proc. IROS 2009*. IEEE/RSJ 2563–2568.
26. Siegel, M., Breazeal, C., Norton, M.I. Persuasive Robotics: The influence of robot gender on human behavior. *In Proc IROS 2009*. IEEE/RSK, 563–2568.
27. Sung, J., Guo, L., Grinter, R.E., Christensen, H.I. My Roomba Is Rambo ": Intimate Home Appliances. *UbiComp 2007*, Springer-Verlag (2007), 145–162.
28. Yamamoto, Y., Sato, M., Hiraki, K., Yamasaki, N., Anzai, Y. A request of the robot: an experiment with the human-robot interactive system HuRIS. *In Proc ROMAN 1992*, IEEE (1992), 204–209.
29. Young, J.E., Sung, J., Voida, A., et al. Evaluating Human-Robot Interaction. *Int J Sol Robotics 3*, 1 (2010).