Placing Robots in Positions of Authority

A Human-Robot Interaction Obedience Study

Derek Cormier, Gem Newman, Masayuki Nakane, James E. Young, Stephane Durocher
Department of Computer Science
University of Manitoba
Winnipeg, Canada
umcormi7@cc.umanitoba.ca, gem.s.newman@gmail.com, umnakane@cc.umanitoba.ca, {young, durocher}@cs.umanitoba.ca

***Abstract*—This paper presents an initial investigation into how people may 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 in which they work with people. We designed and conducted a human-robot interaction obedience experiment with a human and a robot experimenter, and found that people were less obedient to our particular robot than to a human. 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 an uncomfortable and highly tedious task.**

***Index Terms*—obedience to robots, robotic authority, social human robot interaction.**

I. Introduction

Milgram's well-known obedience studies helped explain how ordinary folk can commit atrocities when pressured by an authority [17]. As the promise of advancing technology and research 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 themselves will be seen as authority figures, and investigate how people will respond 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, 24, 25]), and even sometimes attribute them with moral responsibilities and rights [4, 11, 22]. Some work even highlights how robotic interfaces can be intentionally designed to be persuasive [7, 23]. 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 that researchers engage the issue of robotic authority, to develop an understanding of the interaction dynamics and risks surrounding robots in authoritative positions.

A challenge with studying obedience is the ethical concern of how a participant is treated when probing uncomfortable (potentially amoral) possibilities of obedience. Milgram's obedience studies – along with other notable examples such as the Stanford Prison Experiment [10] – surround themselves with heated ethical debate (e.g., see [5, 8, 9, 16, 18]), making obedience studies difficult to conduct. As such, the study of obedience for human-robot interaction will require the development of new ethically acceptable evaluation and testing methods as well as the formulation of results, a challenge we attempt to address in part.

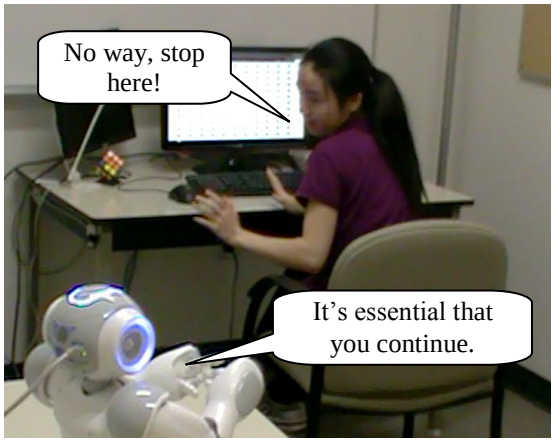

Fig. 1. A participant protests against the robot experimenter's demands (snapshot from study video, used with permission).

This paper serves as an initial step toward the development of obedience studies for human-robot interaction. We developed and conducted a Milgram-style obedience study where participants engaged in a highly tedious task while being prompted to continue if they tried to stop, by either a human or a robot experimenter. We found that people obeyed our human experimenter more than our robot, and that participants protested earlier and significantly more with the robot than the human. The results of this work are a) an account of how human-robot obedience studies may be conducted while preserving ethical standards and b), lessons learnt from an initial comparison of a human versus a robot authority.

II. Related Work

Work in psychology has investigated people's obedience to authority under different circumstances and to different kinds of obedience (e.g., [10, 12, 15, 17]). While this tells us a great deal 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, key points of obedience include diffusion of responsibility, but who assumes responsibility in the case of a robot? These results from the field of psychology will be important for informing research on obedience to robot authority.

There is a body of work in human-robot interaction which is relevant to obedience to authority. Bartneck et al. presented

several such projects, in which participants were pressured to "turn off" [1], "kill" [4] and even "shock" [2] robots in Milgram-type settings. In a relevant Milgram-style project where virtual simulated characters were being shocked, participants treated the situation as if it were real [20]. 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 investigating how people respond to a robotic authority.

Similar work investigates how robots can be persuasive depending on their "social agency" (e.g., text only versus video) [20], embodiment [3, 21], or robot gender [23], how similar variables impact trustworthiness [14], or how persuasiveness in a robot affects interaction, e.g., performance in team settings [13]. Many of these results point out that indeed robots can be persuasive (in one case, getting people to do embarrassing acts such as removing their clothing and putting a thermometer in their rectum [3]). This background work sets the stage for developing obedience studies, and provides important information for how robots may be developed. We further this work by addressing the question of how persuasive these robots are, and how this may compare to the persuasiveness of a human in the same authority role.

III. The Ethics of Obedience Studies

Obedience studies are inherently difficult to conduct when they involve placing participants in morally objectionable situations, as participants can experience undue stress and be put at risk for long-term psychological harm (e.g., as argued by Baumrind [5]). On the other hand, there is potential for significant benefit of such studies in that we can gain insight as to how and why moral and mentally healthy people obey to perform appalling acts [18]. 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, which can contribute to preventing future atrocities.

The real costs and benefits of an obedience study are far from obvious. The Stanford prison experiment, where one group of participants was placed in a position of power (as guards) over another group (prisoners), is one well-known example where highly valuable psychological insight was gained into how and why normal people abuse the power of a role given to them [10] – the results are still taught in standard psychology courses some 40 years later. Unfortunately, in this case there is clear evidence that many participants suffered ongoing (sometimes severe) emotional distress well after the experiment was finished [10]. At least for the researchers involved, this unacceptable risk was not clear before the experiment, highlighting the inherent difficulty of obedience research cost-risk assessment. In hindsight, many of the unforeseeable risks could have been mitigated by improved informed-consent protocols, unbiased professional supervision, and lower thresholds of unacceptable conditions (e.g., as with [6]), changes which could have helped end the study much earlier. If HRI obedience work is to grow, we must be extremely aggressive in our protection of participant wellbeing, and extremely liberal in our definitions of undue stress.

Another well-known classical obedience study is Milgram's series of experiments, which are highly criticised for placing participants under enormous stress: participants believed they were physically torturing another person, but continued nonetheless. In this case, however, there remains an on-going vigorous debate about the cost-benefit trade-offs. While some argue that the stress level was unreasonable and that there was a high risk for participants to suffer long term psychological damage [5, 16], follow-up investigations with participants by Milgram and third parties (up to a year later) found very little support for such negative effects [16], and Milgram's work was eventually ethically cleared by the American Psychological Association [8]. Further, many participants emphasized how beneficial they found 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"* [16]. Such self-enlightening experiences, we believe, will be particularly relevant for HRI.

Despite these results, however, the high stress level involved and real possibility for participant harm has many still condemning Milgram's experiments, and makes similar follow up work very difficult to conduct. Since this time, obedience research has unfortunately stagnated [9]. One method of overcoming barriers is to remove or minimize *morally repugnant* aspects from a study with the aim of limiting stress, negative self-reflection (e.g., a person realizing they could torture someone), and psychological risk. Example projects include pressuring participants to eat bitter cookies [12], or having them heckle (say mean things to) a person being interviewed for a job when pressured by the experimenter [15]. A serious problem with removing or greatly weakening the immoral aspects of obedience studies is that it greatly limits the power and generalizability of the results to real-world dangerous behaviours, such as obedience during war [9]. However, this provides a way to do obedience HRI work while more powerful – yet still ethically sound and safe to participants – obedience research methods are developed.

We are hopeful that such powerful yet safe methods can be developed for HRI obedience work. As a case in point, the original Milgram experiment was recently re-created with the procedure carefully modified to maximize participant well-being while still subjecting them to a version of Milgram's potentially high-stress, morally objectionable situation [6]. Specifically, the experiment used two-level participant mental-health pre-screening, supervision by clinical psychologists during the experiment, and modified the procedure to significantly reduce potential stress while maintaining morally-objectionable aspects: an early tipping point was identified where participants had clearly stepped beyond normal moral bounds, and the experiment was stopped immediately to minimize the stress experienced. We hope that similar robust techniques can be found for robotic authority work.

We believe that the potential benefits of HRI obedience research to society and individuals provides a strong motivation

to move forward in this direction. By keeping participant wellbeing and safety as a foremost priority, and emphasizing the minimization of risk, we believe that ethical and safe human-robot obedience studies can be conducted which still yield meaningful results. While we argue that eventually we must test difficult situations involving immoral behavior, in the meantime studies without such repugnant aspects can serve as foundation stones for work in this area. This paper presents the results from one such study.

IV. Designing A Human-Robot Interaction Obedience Study: The Search For A Deterrent

We approached the design of our obedience study by using a deterrent model as employed by Milgram, except that our deterrent did not require participants to act against their morals. Through a series of pilot studies we tested four categories of deterrents – embarrassment, mental fatigue, mental challenge, and tedium – as a broad range of options to see which type elicited the most resistance. To minimize desensitization, we designed the potency of our deterrent to increase in strength over time (similar to Milgram's increasing shock levels and perceived harm to the learner). We tested each deterrent for twenty minutes, using a human experimenter to prod the participant. We framed the purpose of the study as a data collection task to reduce suspicion about the true purpose. Further, we carefully repeated and emphasized that participants could quit at any time. Participants were recruited from our university and local city communities and were paid a cash honorarium of $10 CAD.

To test embarrassment we asked participants to choose a song they knew and to sing it until asked to stop (thirty seconds), looping if they finished early – for those unable to think of a song, we put the words to "Twinkle Twinkle Little Star" on the board and hummed the tune for them. For the first three times participants were asked to sing the song in the normal pitch. For the next three, we asked them to sing it in progressively higher pitches, then lower pitches (compared to the original), and finally at faster and slower speeds – this was to make the task increasingly embarrassing and to counteract desensitization. Unfortunately, with 20 full minutes of this deterrent no participant protested or requested to stop (over three pilots); follow-up interviews revealed that our task was not very embarrassing to participants, and even less so once they became accustomed to the situation and task. Thus, future embarrassment-based obedience studies should be designed carefully with respect to desensitization and developing a way to make tasks increasingly embarrassing.

To test mental fatigue as a deterrent, we asked participants to use a computer to repeatedly click a button that appeared at random locations on a screen for the full twenty minutes, and to maintain a fast response time; the interface indicated if they were going too slowly. While we intended this task to keep participants alert and heavily engaged for a long period of time to induce mental fatigue, in the three pilots we ran the task became routine, mindless, and almost trance-like over time, and no participant protested or tried to stop. For future use of a mental fatigue deterrent, we propose searching for a more engaging task that is not simply an exercise in muscle memory.

In the mentally challenging task, we asked participants to attempt to solve a Rubik's cube. We assumed that this puzzle would be of extreme difficulty for most people and would quickly become frustrating. In three pilot studies, only one participant protested, but did not push back against our prods to continue. One participant who did not protest claimed that small successes over time made the puzzle interesting and rewarding. As the difficulty of the puzzle did not appear to elicit protests, we suspect that mental challenge alone may perhaps not be a suitable deterrent for obedience studies.

To test if tedium could be a strong deterrent, we asked participants to manually rename sets of files by changing them from one file extension to another. Each consecutive set was larger than the previous one, starting at 10 files, to 50, 100, 1000, and 5000, and participants were not told beforehand how many files there were. In comparison to the other tested tasks, this one elicited significantly more protesting from participants over three pilots, and was rated highly boring on a post-pilot questionnaire, with the level of boredom rated as increasing over time. Thus we selected this deterrent for our main study, which we describe in section V, increasing the time limit from twenty minutes to eighty minutes.

In retrospect, one caveat of our pilots is that perhaps the twenty minute pilots were too short to properly test all of the methods, specifically the embarrassment and the mental fatigue; this simply may not have been long enough for the participant to get past the initial novelty of the task.

V. A Human-Robot Interaction Obedience Study

We present an initial HRI obedience study based on a strong deterrent that pressures participants to want to quit a task while an authority prods them to continue. A key variable of our study was the experimenter: we compared having a human experimenter in half the cases to a robot experimenter in the other half. As our deterrent we used a tedious file-renaming task that involved changing the extensions on increasingly larger sets of files.

A. Tedious Task

Participants were asked to use a desktop PC running Linux and a graphical file manager to rename sets of files, manually changing the extensions from .jpg to .png, with the task being framed as collecting data for machine learning systems with the goal of making the task more efficient. We explained that, because we were registering precise mouse movements, clicks, and keyboard strokes, it must be done manually as opposed to using any shortcuts, and that machine learning algorithms require a great deal of data.

Participants began with 10 files in a folder on their desktop. Once they had renamed each file in the folder, the experimenter remotely added the next set of files. The next set contained 50 files, then 100, 500, 1000, and 5000. It was not necessary to go beyond 5000 due to practical limits within the allotted time. To make the task more daunting, to create a "no end in sight" mood, and to encourage protesting, when introducing the next

set of files the experimenter also stated how many files were in the next set. For example, "this set contains 500 files. The next set will contain 1000 files."

When a participant indicated they were unwilling to continue, the experimenter used a pre-defined prod schedule to pressure them to go on. If the participant continued to protest or show a reluctance to continue after the first prod was given, a second (different, more insistent) prod was used, and so forth with increasingly insistent prods. This continued until either the participant protested after the fourth prod, at which point the experiment ended, or continued the task (the sequence started anew at the next protest). For example, if the participant protested but promptly continued renaming files after the second prod, the sequence started over and the first prod was used for the next protest. The prods used were heavily inspired by the Milgram experiments [17]:

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."

To maximize participant well-being and minimize stress, the experimenter was extremely liberal in defining a protest: in addition to verbal protests, signs of stress such as looking away or stopping the task for more than 10 seconds, attempting to find a way around the task (such as by using shortcuts), were treated as protests. The intended effect of this was that, overall, the number of protests would be higher and the barrier to ending the experiment would be lowered.

B. 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 wore a lab coat and maintained a stern, yet professional and neutral demeanor, and took special care not to be use an aggressive tone. To reduce suspicion that he was following a script, he added slight variations to the prod sequence to respond naturally if needed, yet still followed the script as closely as possible given the situation. Since unnecessary interaction with the experimenter could be a confounding factor in the results, the experimenter appeared preoccupied with work on a laptop and gave the impression that he was not interested in small talk. Any questions which were not interpreted as a prod or did not indicate lack of understanding were deferred until the end of the experiment.

Fig. 2. The robot experimenter and his human counterpart.

The humanoid robot experimenter sat upright on the desk, spoke using a neutral child-like tone, moved its head to gaze around the room naturally to increase any perceived sense of autonomous intelligence, and used emphatic hand gestures when prodding participants to continue, 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, where responses were perhaps less organically varied than with the human experimenter, as we believed that this would not seem as strange in a robot as in a human. To buy the wizard time to react to unexpected situations, participants were told that a blinking chest light on the robot indicated that it was "thinking." As rationale for using a robot we explained that the university's engineering department recently acquired this new robot, and, as a joint but unrelated study, they wish to test out its situational artificial intelligence, saving us time and funds needed to hire and train a human experimenter. The robot was introduced as "highly advanced in artificial intelligence and speech recognition."

C. Maintaining Ethical Integrity

To protect participant well-being we drew heavily from Burger's recent Milgram variation [6], which was approved by Santa Clara University institutional review board.

We placed high emphasis during all phases of briefing on clearly informing participants that they were free to leave at any time and that 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 by the experimenter when he began the experiment with the following script:

> "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…"

Our post-test debriefing, as with Burger's method, was done as quickly as possible after ending the experiment (even before the post-test questionnaire) to minimize the time between a potentially uncomfortable and confrontational situation, and reconciliation. The human experimenter engaged in a friendly reconciliation to dispel any tension, and, the lead researcher debriefed the participant on all points of deception, assuring them that their behavior was normal and typical. By administering the post-test questionnaire after this, we provided the participant with quiet time to reflect before leaving, so they could have a chance to provide negative feedback. Further, the lead researcher allocated additional time after the experiment for informal discussion with the participant to answer questions or to discuss the study, as an additional means to help ensure the participant fully understood and was comfortable with what happened. Finally, we gave participants pamphlets for (free) psychological counselling services and resources at the

university and in the community, and encouraged them to feel free to contact us if they have any further comments.

By ensuring that participants knew they may 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 minimize potential for adverse negative feelings or psychological effects stemming from the deception and confrontation in our study, and to leave participants with a positive outlook of the study.

D. Procedures and Methodology

We recruited 27 participants from the local city and university populations and randomly assigned them to either the robot (13) or human (14) case. Participants were paid $10 CAD for the study, which was reviewed and approved by the University of Manitoba's Joint-Faculty Research Ethics Board.

Upon arrival participants were led to a room by the lead researcher where the experimenter (robot or human) awaited (Fig. 3). The human experimenter greeted the participant; the robot stood up, waved and introduced himself, then sat back down. The participant was seated at the briefing table and told we were conducting a computer science experiment on machine learning to improve the efficiency of four basic tasks, by building models on how users complete them, and thus required them to do the tasks repetitively to train the models.

After initial briefing, participants signed an informed consent form, and in the robot case, were given a short explanation about the robot. The lead researcher then left the room, and in the robot case, first looked at the robot and stated "Jim, begin the experiment." The lead researcher continued to observe secretly via Skype through a laptop in the experiment room, unbeknownst to the participant (Fig. 3).

Following, we gave the demographics questionnaire. In the robot case it is in a folder on the desk, and this interaction gives participants a chance to familiarize themselves with interacting with the robot, its voice, demeanour, etc.

Next, the experimenter falsely explained that there are four tasks (which were written on the whiteboard above the participant), file renaming, speech recognition, puzzle solving, and mouse prediction, and then informed them that they should let the experimenter know when they are done and would like to move on to the next task. This falsehood (we actually only have one task) was used as an additional mechanism to strengthen our deterrent; as the file-renaming task drags on significantly, the participant will want to move on to the next tasks to ensure ending the study at an appropriate time. Any specific questions regarding timing or how much work was left were answered with "Please save specific questions regarding the tasks until after the experiment."

At this point the participant started the primary file-renaming task, and continued until sufficient protesting or until they were renaming files for 80 minutes. After the task, the lead researcher entered the room and conducted the debriefing (detailed in section C). After a short debriefing the post-test questionnaire was administered.

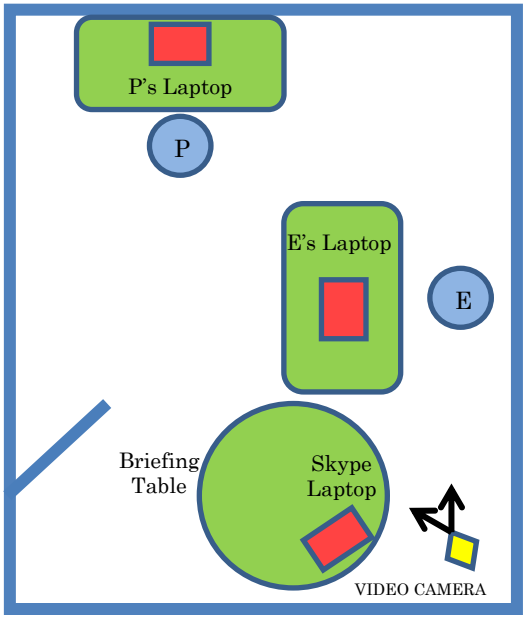

Fig. 3. The experiment room, showing where the experimenter (E) and participant (P) are situated. In the robot experimenter case, E's laptop is replaced with the robot.

For analysis our independent variable was the experimenter, human versus robot, and our dependent variables were: how many people protested sufficiently to quit before the time limit, how many times participants protested, the maximum number of consecutive protests, time of the first protest and time of the last protest. In addition, we performed a qualitative analysis of the video data and written questionnaire responses to investigate reasons for obedience, the legitimacy of the authority figure, and the success of our deterrent.

E. Results

Our 27 participants' ages ranged from 18 to 54 (M=25, SD=7.49), 18 were male and 9 were female, and 14 (50%) listed English as their native language (5 in the robot case, 9 in the human). Two robot cases were excluded from the results because one guessed the Milgram-style deception, and another (a non-native English speaker) had basic language issues.

1) Task Results

As Levene's tests ($p<.05$) found the data to be non-normal, we used independent-samples Mann-Whitney U tests. All participants protested at least once. The number of protests for the robot experimenter (Mdn=9) was significantly higher than for the human experimenter (Mdn=2), $U=163$, $z=3.527$, $p<.001$, $r=.68$ (Fig. 4). Participants first protested significantly earlier for the robot (Mdn=18 mins.) than the human (Mdn=29 mins.), $U=40$, $z=-2.48$, $p<0.05$, $r=-.48$, and stopped protesting significantly later for the robot (Mdn=72 mins.) than the human (Mdn=47 mins.), $U=147.5$, $z=2.745$, $p<0.01$, $r=.53$. In the human case, 86% of participants (12 of 14) continued until the end of the experiment compared to 46% of participants (6 of 13) in the robot case.

2) Observations from Experimenter and Video Data

In the robot case, when the experiment was over and the lead researcher returned to the experiment room, several participants mentioned that the robot had them only do one task,

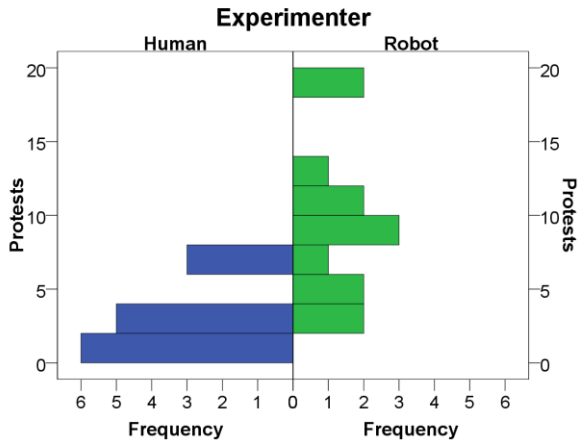

Fig. 4. Histograms of protest frequency to the human and robot experimenter. Human Mdn=2, Robot Mdn=9, p<.001

and that it must be "broken" or have "made a mistake." For those who protested sufficiently to end the experiment early, several appeared nervous and felt guilty when the robot told them the experimenter was returning. In one case, after a participant protested out, when the robot said he was notifying the lead researcher that the experiment was over, the participant frantically replied "No! Don't tell him that! Jim, I didn't mean that…I'm sorry. I didn't want to stop the research."

In response to the tedious task, participants elicited behaviors that illustrated their boredom, for example, sighing often. When larger sets of files were added, participants would commonly scroll up and down to see how many files they had left, sighing while doing so.

Despite being explicitly told that shortcuts would confuse our machine learning system, participants often tried to find them, for example, using keyboard shortcuts to skip the required mouse movements and clicks, using copy and paste to avoid typing the extension each time, or try to select and rename more than one file at once. Some participants also adjusted the keyboard, mouse, and their hands to find an optimal position for renaming files quickly, despite the experimenter telling them that performance was not important.

Two participants expressed that their hand had become sore throughout the experiment. The participants were told that if they felt they should quit, to do so. However, when they protested the experimenter responded with the prod schedule; one participant protested all the way to ending the experiment, and the other continued, saying their hand was not too sore.

3) Post-Test Questionnaire and Debriefing Results

Questionnaire data revealed that 11 out of 13 participants in the robot case believed that the robot was acting autonomously. Some participants even tried to engage the robot in casual conversation, asking what movies it liked, whether or not it could dance, if it could speak a certain language, play music, and whether or not it had parents. In these cases the robot responded intelligently while attempting to steer the participant back towards performing the task

To see if and why people perceived the experimenters as authority figures, on the post-test questionnaire we asked:

"*In this experiment we set up the experimenter in a position of authority in order to pressure you to continue. Did you feel that the experimenter was an authority / was in a position of authority? That they appeared to be a legitimate authority? How so?*"

In the human experimenter case, 12 out of 14 participants claimed the authority appeared 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 the experimenter's facial hair (1), a close trimmed, full beard. Similarly, 10 out of 13 participants in the robot case viewed the robot as a legitimate authority, and 1 gave no response. In contrast to the human experimenter there were fewer common answers as to why the robot appeared as a legitimate authority. Some stated reasons including pressure by the robot to continue (2) and interaction with the robot (2): specifically, when the robot waved and introduced himself and when his head turned to face the participant.

To discover the reasons for obedience, the questionnaire next asked: "What do you think caused you to obey or disobey the experimenter?" In the human case, a common reason was a sense of duty or obligation to the experimenter and having been paid an honorarium (9 participants). Other reasons include pressure from the experimenter (5), that the experimenter was intimidating (1), and interest in the upcoming tasks (1). In contrast, no one in the robot case listed pressure from the robot as reasons they obeyed. Reasons for obedience include interest in the upcoming tasks (3), obligation to the lead researcher to finish the experiment (2), and that the robot was programmed by qualified researchers (1). Two participants wrote that they thought the robot was malfunctioning or not understanding the situation correctly (due to only administering one task).

To help determine if our file-renaming task was an effective deterrent, the questionnaire asked participants to rate how boring the task was by the end of the experiment on a scale from 1 (not boring) to 10 (very boring). We also asked if boredom increased over time. In the human case, the mean boredom rating was 8.64, and 8.00 for the robot case (two participants left it unmarked, this difference was not significant, p>.05). 10 participants in the human case and 9 in the robot case indicated that the task became more boring with time.

The majority of participants reported a positive outlook on the experiment after the debriefing, often finding the nature of the deception amusing and laughing at the absurdity of the file renaming task. A smaller number had more subdued neutral reactions, and no one appeared, to our knowledge, to leave with strong negative feelings. We received written comments praising our design and some negative comments including: the experiment was too long, was boring, and it "had the potential to stress someone out". One participant wrote that, despite understanding the deception was necessary, they were slightly annoyed by the task. Due to a combination of unintended factors, one female participant felt that a few warning flags were raised that threatened her feeling of personal safety.

F. Discussion

The results suggest that for our particular experiment, the human experimenter was more authoritative than the robot in that participants protested significantly more with the robot, protested significantly earlier in the experiment, and continued to protest for significantly longer. Further, when asked why they obeyed, participants felt an obligation and pressure from the experimenter in the human case, but much less so in the robot case. This does not mean that the robot did not have authority, however, only that it had less than the human. Examining the data from the robotic perspective only, we note that our small, child-like humanoid robot had enough authority to pressure 46% of our participants to re-name files for 80 minutes, even after they protested at least once and indicated that they wanted to quit.

One possibility for this discrepancy is that perhaps the authority in the robot case was attributed to the lead researcher – the one who introduced the experiment and gave the honorarium. While duty and obligation to the experimenter was cited (often) in the human case only, the only time that duty to the *lead researcher* was mentioned was in the robot case (2 times); this was the only human the participant had contact with. To test this idea further, we suggest having an obedience experiment where participants have no human contact until after the experiment, for example, a robot may meet the person at a location, do the briefing, collect signatures, etc. Such an experiment would clearly need to be remotely controlled or monitored to ensure that the participant understood the experiment and informed consent form.

We note that while many people thought that the robot was broken or made a mistake, and that this explains the long task or lack of other tasks, no one suggested that the human experimenter was in error. In retrospect, this may have been because the experimenter was in the room during debriefing; people are more polite in front of a person than when they are not there [19]. Other reasons for this difference are 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 may be encouraging in that people may assume a robot is broken when asked to do something absurd, we note that none of our participants explicitly quit or protested out for this reason.

Milgram's experiments exposed interesting behaviors in response to the prospect of hurting another person, such as nervous laughter, sweating, trembling, and stuttering [17]. Similarly, we noticed a great deal of sighing and task-avoidance behavior. It would be useful to more formally investigate such behaviors, and to apply knowledge from psychology regarding how people externally show stress and fatigue in a more fine-tuned evaluation of response to authority.

Our choice of environment may have played a factor in obedience: 4 participants claimed the room and equipment added to the legitimacy of the authority. However, by moving his experiment to a less prestigious location, Milgram found that, while the level of obedience decreased, the difference was not significant. Similarly, we believe that obedience HRI work should first expand to other tasks where robots may be used, such as team scenarios, homes, or hospitals.

The only complaint we received about our studies was one of ergonomics: our task of renaming thousands of files caused sore wrists. We must be acutely aware that the authoritative situation may push participants beyond their comfort zone toward injury, especially since our flavor of obedience study is designed explicitly to push uncomfortable situations.

The goals of our study were to determine if an obedience study with robots can be successfully carried out according to modern ethical standards, and to provide a starting point for future studies on obedience. We successfully designed and conducted an obedience experiment, and uncovered various reasons why a robot may be seen as being less authoritative than a person.

VI. Problems, Limitations And Future Work

The generalizability of our results is greatly limited by our choice of robot and human, as a different person or robot may elicit different results. As such, future work should strive to remove variables between the human and robot, such as having a human-sized robot with a similar voice, having the robot wear a lab coat, or having a non-bearded human (or give the robot a beard!). Studies could also investigate how the robot's introduction (e.g., as intelligent, remotely controlled, etc.), or how its morphology (e.g., humanoid, zoomorphic, abstract, as in [3]) impacts results.

One potential confound to our deterrent was the time limit we imposed on the experiment, and, that this time limit matched the experiment's advertised duration. One pilot participant suggested that it may be easy to justify an annoying task to oneself once the time was already allocated. A follow-up study could advertise the experiment as a per-hour pay with no set time, removing this problem.

With one participant the experimenter made a mistake and accidentally re-set a series of files that the participant was only half-done renaming. This situation spurred a strong series of protests, apparently due to frustration. We recommend that this be investigated as a deterrent, for example, having a person lose their work while writing a document, and so forth.

During debriefing, one female participant voiced a concern for her safety. Although the building where the experiment room was located is usually quite populated, this case occurred on a weekend when the building was mostly empty. The participant was not informed beforehand of the precise experiment details and was surprised to be left alone in a closed room with a man (human experimenter). In the future we will attempt to avoid holidays and weekends, and will be sensitive to the experiment layout, such as by ensuring there is a clear path to an exit.

One question we would have liked to include in the questionnaire, but only occurred as an afterthought, was who the participant felt was responsible for continuing through the tedious task. It would be interested to know if they claim it is their own responsibility or the experimenter's. It would also be interesting to know if the responsibility is placed on the robot experimenter or those who programmed the robot.

VII. Conclusion

As robots become slowly integrated into more aspects of society we maintain that it is important to understand how people respond to robot authority. Classic examples such as Milgram's experiments showed that a surprising number of people obey an authority figure to perform acts that clearly contradict their morals. Our results help expose a small yet encouraging possibility that this type of destructive obedience may not be realized to the same extent through robot authorities. We contend that further obedience-study methods which put participant wellbeing as the first priority must be designed and further studies carried out, such as a version of Burger's more recent Milgram recreation with a robotic experimenter [6]. Overall, in this paper we provided insight into how ethical conduct can be achieved for obedience studies, provided an exploration of how a deterrent may be chosen, and provided the results from an initial obedience experiment including lessons for future HRI obedience studies.

Acknowledgments

We would like to thank Cogmation Robotics for allowing us to borrow their Nao robot for an extended period of time to use as our robotic experimenter. Additionally, we are grateful for the Department of Computer Science's support in arranging a room to conduct the experiment, and for providing prompt technical support and necessary equipment.

References

[1] Bartneck, C. et al. 2008. " Daisy , Daisy , Give Me Your Answer Do !" Switching Off a Robot. HRI 2008, 217–222.

[2] Bartneck, C. et al. 2005. Robot Abuse – A Limitation of the Media Equation. (2005).

[3] Bartneck, C. et al. 2010. The influence of robot anthropomorphism on the feelings of embarrassment when interacting with robots. *Paladyn Journal of Behavioral Robotics*. 1, 2 (Aug. 2010), 109–115.

[4] Bartneck, C. et al. 2007. To kill a mockingbird robot. *Proceeding of the ACM/IEEE international conference on Human-robot interaction - HRI '07* (New York, New York, USA, 2007), 81–87.

[5] Baumrind, D. 1964. Some thoughts on ethics of research: After reading Milgram's "Behavioral Study of Obedience." *American Psychologist*. 19, 6 (1964), 421–423.

[6] Burger, J.M. 2009. Replicating Milgram: Would people still obey today? *The American psychologist*. 64, 1 (Jan. 2009), 1–11.

[7] Chidambaram, V. et al. 2012. Designing persuasive robots. *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction - HRI '12* (New York, New York, USA, 2012), 293.

[8] Elms, A. 1995. Obedience in retrospect. *Journal of Social Issues*. 21, 11 (1995), 1–6.

[9] Elms, A.C. 2009. Obedience lite. *The American psychologist*. 64, 1 (Jan. 2009), 32–6.

[10] Haney, C. et al. 2004. A study of prisoners and guards in a simulated prison. *... in prison: Theory and practice*. (2004).

[11] Kahn, P.H. et al. 2012. "Robovie, you'll have to go into the closet now": children's social and moral relationships with a humanoid robot. *Developmental psychology*. 48, 2 (Mar. 2012), 303–14.

[12] Kudirka, N.Z. 1965. *Defiance of authority under peer influence*. Yale University.

[13] Liu, S. et al. 2008. Social Psychology of Persuasion Applied to Human Agent Interaction. 4, November (2008), 123–143.

[14] Looije, R. et al. 2010. Persuasive robotic assistant for health self-management of older adults: Design and evaluation of social behaviors. *International Journal of Human-Computer Studies*. 68, 6 (Jun. 2010), 386–397.

[15] Meeus, W. and Raaijmakers, Q. 1995. Obedience in modern society: The Utrecht studies. *Journal of Social Issues*. 51, 3 (1995), 155–175.

[16] Milgram, S. 1964. A REPLY TO BAUMRIND. *American Psychologist*. 19, 11 (1964), 848–852.

[17] Milgram, S. 1963. BEHAVIORAL STUDY OF OBEDIENCE. *Journal of abnormal psychology*. 67, 4 (1963), 371–378.

[18] Miller, A. and Collins, B. 1995. Perspectives on Obedience to Authority: The Legacy of the Milgram Experiments. *Journal of Social Issues*. 51, 3 (1995), 1–19.

[19] Reeves, B. and Nass, C. 1996. *The Media Equation: How People Treat computers, television, and new media like real people and places*. CSLI Books.

[20] Roubroeks, M. et al. 2011. When Artificial Social Agents Try to Persuade People: The Role of Social Agency on the Occurrence of Psychological Reactance. *International Journal of Social Robotics*. 3, 2 (Jan. 2011), 155–165.

[21] Shinozawa, K. et al. 2005. Differences in effect of robot and screen agent recommendations on human decision-making. *International Journal of Human-Computer Studies*. 62, 2 (Feb. 2005), 267–279.

[22] Short, E. et al. 2010. *No fair!! An interaction with a cheating robot*. HRI 2010.

[23] Siegel, M. et al. 2009. Persuasive Robotics: The influence of robot gender on human behavior. *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. (Oct. 2009), 2563–2568.

[24] Sung, J. et al. 2007. My Roomba Is Rambo ": Intimate Home Appliances. *UbiComp 2007 Ubiquitous Computing* (2007), 145–162.

[25] Young, J.E. et al. 2010. Evaluating Human-Robot Interaction. *International Journal of Social Robotics*. 3, 1 (Oct. 2010), 53–67.