Please Continue, We Need More Data: An Exploration of Obedience to Robots

Denise Y. Geiskkovitch[1†], Derek Cormier[3], Stela H. Seo[2], James E. Young[2]

[1]Department of Psychology
[2]Department of Computer Science
University of Manitoba
Winnipeg, Manitoba
Canada

[3]Department of Computer Science
University of British Columbia
Vancouver, British Columbia
Canada

We investigated obedience to an authoritative robot and asked experiment participants to do a task they would rather not do. Obedience questions are increasingly important as robots participate in tasks where they give people directions. We conducted a series of HRI obedience experiments, comparing a robotic authority to other authority instances, including: (*i*) a human, (*ii*) a remote-controlled robot, and (*iii*) robots of variant embodiments. The results suggest that half of participants will continue to perform a tedious task under the direction of a robot, even after expressing desire to stop. Further, we failed to find an effect of robot embodiment and perceived autonomy on obedience. Instead, the robot's perceived authority status may be more strongly correlated to obedience.

Keywords: human-robot interaction, obedience, persuasion.

1. Introduction

Milgram's well-known obedience studies help explain how ordinary people can commit atrocities when pressured by an authority (Milgram, 1963). While not generally thought of as authority figures, robots are becoming more common in households, schools, hospitals, and disaster sites, and individuals often respond to them as social entities, sometimes even attributing them with moral responsibilities and rights (Bartneck, Verbunt, Mubin, & Al Mahmud, 2007; Kahn et al., 2012; Short, Hart, Vu, & Scassellati, 2010)—as such, it is important to investigate how people will likewise respond to these social robots being authority figures and how robot design may impact such a response.

It is already well established that people tend to anthropomorphize robots and treat them as social entities (e.g., Bartneck, Verbunt, et al., 2007; Sung, Guo, Grinter, & Christensen, 2007; Young et al., 2011). Some research highlights how robotic interfaces can be intentionally designed to be persuasive (Chidambaram, Chiang, & Mutlu, 2012; Siegel, Breazeal, & Norton, 2009). Chidambaram et al. (2012) tested how combinations of verbal and nonverbal cues provided by a robot affected participants' compliance. They found that participants were more often persuaded by the robot when it used nonverbal cues (either vocal, bodily, or a combination of the two), than when it did not provide those cues. In either case, the robot was able to persuade individuals to modify their answers. Siegel et al., on the other hand, studied how a robot's gender might impact participants' donations. Participants interacted with either a male or female robot that asked them to donate money. Men were found to be more likely to donate money when they had interacted with a female robot, but the robot's gender did not affect women's donation behavior. Save for a small number of tangentially related studies, little is known about how people react to robots in authority positions. Moving forward in the field of human-robot interaction (HRI), we argue that it is crucial to engage the issue of robotic authority and to develop an understanding of interaction dynamics and risks surrounding the placement of robots in potentially authoritative positions.

A prohibiting challenge with studying obedience has been the ethical concern of how participants are treated when probing uncomfortable (potentially amoral) possibilities. Milgram's (1963) obedience studies—along with

other notable examples such as the Stanford Prison Experiment (Haney, Banks, & Zimbardo, 1973)—are surrounded by ethical debate (e.g., Baumrind, 1964; Elms, 2009; Elms, 1995; Milgram, 1964; Miller & Collins, 1995), while similar studies are difficult to conduct. The study of obedience for HRI requires the development of new ethically-acceptable evaluation and testing methods, a challenge we begin to address in this paper.

Although obedience, in general, is an established area of psychology, much less is known about how people react to being pressured by robots in authority positions. This paper serves as a step in this direction: We developed an original HRI obedience study design that adheres to modern ethical standards, with a design that prods participants to continue a task they would rather stop doing, and conducted an experiment with several instance variants to explore how aspects of the robot and scenario may impact obedience. In particular, we compared participant obedience to an autonomous humanoid robot versus obedience to a human, as a base case to validate our study design and provide a baseline of potential expected obedience in the given scenario. We also compared an autonomous humanoid robot case to a non-autonomous, remote-controlled humanoid instance to test how perceived autonomy (versus being a proxy) may influence obedience. Finally, we compared the autonomous humanoid robot to different autonomous embodiments (non-humanoid disc-shaped robot and a non-robotic computer server) to investigate the impact of morphology on obedience.

The primary research question behind our experiments was as follows: Will people obey a robot placed in a position of authority that tells them to continue a task after they have indicated they want to quit? Rather than defining an expected level of obedience, our aim was to establish an initial baseline of whether or not any participants would obey, and if so, how many. Further, we aimed to provide insight into the interaction, such as how long participants obeyed, and how they disobeyed (e.g., what verbal or non-verbal strategies they employed).

Our secondary research questions involved learning more about what factors might influence obedience. One question was whether the perception of robot autonomy (or the lack thereof) would influence obedience. Our hypothesis was that participants would obey a remote-controlled robot more than an autonomous one, because they would be obeying a human via the robot as a proxy, and not a robot per se; we assumed that it is more natural to obey a human authority than an artificial one. Our final research question was whether the robot's embodiment affected obedience, with the embodiment on a continuum from human-like to machine (i.e., humanoid, disc-shaped robot, or computer server). We hypothesized that human-like robots would elicit more obedience than machine-like robots, as per our assumption above, that people are more likely to obey a human (or human-like robot) than a more conventional machine.

In our experiment, nearly half of people obeyed the robots to continue a highly tedious task until the end, despite repeatedly requesting to quit the experiment; many participants (including those who ultimately quit) argued and rationalized with the robot in their attempts to quit, even if they continued the experiment. As a baseline, we did find that more participants (86%) obeyed a human experimenter than an autonomous humanoid robot (45%) to continue a highly tedious task. While we did not find an effect of robot autonomy or embodiment on obedience in the remaining conditions, we found evidence for a relationship between participants' perception of the robot as an authority figure and their obedience toward it.

The contributions of this research are as follows:

a) an HRI-specific analysis of obedience and recommendations for ethically conducting obedience HRI studies that protect participant well-being;
b) an original HRI obedience study design that protects participant well-being and maintains ethical standards;
c) evidence that participants may obey robots to continue doing things that they do not want to do, additionally validated with a study replication and across conditions;
d) initial exploration into the effects of robot autonomy and embodiment on obedience to robots, an initial indication that these may not be significant factors to obedience, and recommendations for conducting similar studies;
e) directions for future obedience work grounded in our experimental results.

Elements of early results from this work have appeared previously (Cormier, Newman, Nakane, Young, & Durocher, 2013) and include the following: early discussions on obedience and ethics, an initial obedience study design, and initial pilot study results. This paper summarizes previous work and details results in the respective sections below, extends the analysis, and presents all study conditions that are integral to a new, overarching discussion. Our work serves as a single piece of a broader, ongoing HRI obedience research agenda. We believe that these initial results will ultimately inform the research and design of robots that could be placed in authoritative (and potentially destructive) situations (e.g., the military), and we encourage research that will lead to robot design that

attempts to prevent potentially harmful obedience. Generalizing this work directly to instances where people may conduct harmful or morally repugnant acts (as in the Milgram experiments) is difficult; however, we believe that this exploration of obedience in the face of a deterrent is an important first step in this direction.

2. Related Work and Background

Obedience to authority has been investigated in psychology under various conditions (e.g., Haney et al., 1973; Meeus & Raaijmakers, 1995; Milgram, 1963), providing insight into how people respond to authority figures. Robots, however, provide a unique interaction experience, so it may not be appropriate to directly generalize psychology results to robots; robots are machines, not people, but have a life-like agency, (active agency; Young et al., 2011), and people treat them in many ways like living things. Because of this, it is uncertain whether people will treat the robot like a machine or a human in obedience scenarios. For example, a key point of obedience includes diffusion of responsibility, but would a person deflect responsibility for their actions to an autonomous robot? Or perhaps to an individual perceived as being above (in authority) or behind (controlling) the robot (e.g., owner, company, government, etc.)? While psychology provides insight into such questions, obedience research needs to be explicitly reconsidered in the context of robots in authority positions.

The field of HRI has a burgeoning body of work along the lines of obedience research. Much of it falls under the umbrella of persuasive robots, exploring, for example, how a robot's embodiment (Bartneck, Bleeker, Bun, Fens, & Riet, 2010; Roubroeks, Ham, & Midden, 2011; Shinozawa, Naya, Yamato, & Kogure, 2005) or designed gender (Siegel et al., 2009) can impact a person's trust (Looije, Neerincx, & Cnossen, 2010), or how persuasive the robot is. Others have investigated how a persuasive robot character can affect interactions, including performance in team settings (Liu, Helfenstein, & Wahlstedt, 2008). A pilot study also compared requests from an unfamiliar robot and a familiar human experimenter (Yamamoto, Sato, Hiraki, Yamasaki, & Anzai, 1992). This body of work suggests that robots can in fact be persuasive; the robot, however, is not usually placed into an explicit authority position, and the work does not include obedience-style deterrents (uncomfortable elements), which help investigate if a robot can explicitly pressure a person to do something they would rather not do.

Along persuasion lines, robots can help pressure people to do things, such as completing exercises in rehabilitation therapy (Matarić, Tapus, Winstein, & Eriksson, 2009), reducing their energy consumption (Ham & Midden, 2014), and maintaining their fitness goals—one study found that robotic progress-tracking systems can be more effective than simple computer or pen and paper methods (Kidd, 2008). Such work highlights potential positive applications of obedience work, where robots can persuade and encourage people to attain their goals. An important element of this work is that the individual being pressured wants the pressure and has a desire to realize a goal; conversely, in the more general obedience case that we study, the participant may strongly object to what they are being pressured to do, and yet through pressure may do it regardless.

Bartneck and colleagues conducted a series of related studies where participants were pressured by a human (not a robot) to "turn off" (Bartneck, van der Hoek, Mubin, & Al Mahmud, 2007) or "kill" (Bartneck, Verbunt, et al., 2007) robots. Another study (Bartneck, Rosalia, Menges, & Deckers, 2005) had people administer shocks to robotic or virtual entities to replicate the Milgram (1963) experiments. Bartneck et al. (2005) had participants teach a robot a set of words and administer a shock to the robot if it made a mistake. The robot would then utter complaints similar to those in Milgram's experiment. Here, the research investigated whether participants resisted harming non-living entities, and why they resisted if they did. 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 an individual to do something they would rather not do.

Perhaps the most relevant prior work is a case where various robots pressured participants to do embarrassing acts (an indirect deterrent). The results were striking: In the context of a fake medical examination, a robot was able to persuade participants to remove their clothing and put a thermometer in their rectum while on camera (Bartneck et al., 2010). Such results motivate the importance of improving our understanding of how robots can have authority over people, and to what extent people will obey robots.

While the body of background work described above emphasizes the importance of researching robotic authority, the lack of examples and results specifically pertaining to robotic authorities pressuring people in the face of a deterrent highlights the limited knowledge regarding people's obedience to robots. Our research complements and extends this body of work and has an important original angle: We use explicit pressure and direct demands in the face of a deterrent instead of motivation, subtle persuasion, or indirect deterrents. In this paper, we specifically investigate how participants react in these direct-deterrent conditions, and how such obedience may be impacted by perceptions of the robot's autonomy or embodiment.

3. Ethics of Obedience Studies

As discussed in our prior work (Cormier et al., 2013), obedience studies are inherently difficult to conduct when they involve placing participants in objectionable situations. These studies can be helpful when investigating if robots can push people to do bad things, for example, to help reduce the likelihood of events such as those that occurred in the Stanford Prison and Milgram Experiments (Haney et al., 1973; Milgram, 1963) from taking place with an authoritative robot in charge. This difficulty arises 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; Baumrind, 1964). On the other hand, there is potential for significant benefit from such studies in that we can gain insight into how and why moral, mentally healthy individuals obey authorities to do appalling acts (Miller & Collins, 1995). For obedience work with robots, it is important to balance the potential risks to participants with the potential for improved understanding of how, when, and why people may obey robots. We therefore use this section to outline the struggles that previous research has found in maintaining ethics in obedience studies and our motivation and thinking process in establishing a safe paradigm for our studies.

The balance between 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 healthy people may abuse power (Haney et al., 1973)—the results are still taught in psychology courses 40 years later. However, many participants suffered (sometimes severe and ongoing) emotional distress (Haney et al., 1973). These risks were not obvious to the researchers beforehand, highlighting the difficulty of risk-benefit assessment. In hindsight, some risks could have been mitigated by having improved informed-consent protocols, unbiased professional supervision, and lower thresholds of unacceptable conditions (e.g., Burger, 2009). If HRI obedience work is to grow, we need to accept benefit-risk assessment difficulties and must be aggressive in our protection of participant well-being.

Milgram (1963) performed a series of experiments where participants believed they were physically harming another person under an authority figure's direction. They were instructed to administer increasingly strong shocks to the learner (a confederate actor) in an adjacent room and continued to do so under pressure of the experimenter. Despite the learner screaming in pain and eventually going silent (ostensibly due to a heart attack), and despite participants' agitation at the unpleasant task, 65% of participants continued on to the final shock level. The experiment highlighted that people may cross their 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 still an ongoing vigorous debate about the risks and benefits. While some argue that the unacceptable stress level created risk of long-term psychological harm (Baumrind, 1964; Milgram, 1964), little support for negative effects was found in follow-up investigation (Milgram, 1964), and the experiment was eventually ethically cleared by the American Psychological Association (Elms, 1995). 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" (Milgram, 1964). If risks can be managed, creating enlightening experiences with robots will be important for HRI's future.

Even the minor possibility for participant harm has many still condemning such work, and obedience research has stagnated (Elms, 2009). 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 (Kudirka, 1965) or to heckle (say mean things to) an interviewee (Meeus & Raaijmakers, 1995). While weakening moral aspects greatly limits the generalizability of results to real-world dangerous behaviors (Elms, 2009), it provides a way to conduct obedience HRI work while more powerful—yet still ethically sound and safe to participants—obedience research methods are being developed.

Toward more ethically sound research methods in obedience, 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 (Burger, 2009); a tipping point was identified where participants had very clearly stepped beyond normal moral bounds but precluded the highest levels of potential stress. Additionally, 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 the study of robot obedience. Progress will require safe and ethical evaluation methods development that considers participant well-being as a foremost priority, while containing elements of obedience and pressure such that results are meaningful and applicable to real-world situations. Our work provides one such study design in this direction and associated results.

4. HRI Obedience Study Design

Our approach to study design was to use a deterrent to encourage participants to want to quit, while having a robot prod them to continue (inspired by the Milgram experiments; Milgram, 1963). 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 exploration), testing if participants 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, and each participant completed all tasks below. The participant was said to protest if they stopped completing the task for 10 seconds or if they complained to the experimenter.

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

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

Participants were also asked to solve a Rubik's Cube. We believed that after initial successes (e.g., one color solved) the puzzle would quickly become too difficult, participants would want to stop, and the task would serve as an intellectually challenging deterrent. There was only one protest during the task, and results indicated that participants generally 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, as described below. More details on these pilot studies were published previously (Cormier et al., 2013).

Overall, we faced various challenges with finding an effective deterrent, mainly: (*i*) that desensitization can quickly weaken a deterrent, as can embarrassment, (*ii*) that repetitive behaviors can become trance-like, and (*iii*) that 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 in some cases.

4.1 Maintaining Ethical Integrity

As previously mentioned, in the past, the ethics of conducting obedience studies has been heavily questioned. With this in mind, one of our main goals was to ensure the ethical integrity of our study paradigm to protect the participants. Our study design drew heavily from Burger's (2009) Milgram experiment variation. We clearly emphasized that participants were 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 carried-out as quickly as possible after the task (even before the post-test questionnaire). The human researcher deliberately engaged in a friendly demeanour (to dispel tension) and debriefed the participant on all points of deception. To counteract embarrassment at not realizing the real intent, we assured participants 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 researcher 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 had troubled them, and we 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 their participation in the study and to minimize potential for adverse negative psychological effects stemming from the deception and confrontation used.

4.2 Tedious Task

Participants used a desktop PC file manager (GNOME 3 Fallback Mode) to manually rename file sets from "jpg" to "png" extensions. To avoid suspicion about the experiment's purpose, we framed the task's goal 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 (such as keyboard copy and paste) must be avoided, to prevent participants from bypassing the tedium.

The task began with 10 files to rename. Upon completion, a new (increasingly larger) set was given to the participants (through remote networked control) of the sizes: 50, 100, 500, and 1000 files. Participants were not informed of file counts beforehand, in order to increase task tedium (i.e., the feeling that there is no end) but were told the size of the present files set and the following set. For example, the experimenter would say, "This set contains 50 files, the next set will contain 100 files" to encourage them to want to quit. The increasing amount of files in subsequent sets was important, because participants may have become desensitized to the amount of files being added if that amount remains the same from one set to another. In addition, although there was only a single task (file renaming), participants were told that there were four separate tasks (file renaming, speech recognition, puzzle solving, and mouse prediction) and were instructed to let the experimenter know when they wanted to move on to the next one. This falsehood was intended to pressure participants to want to switch tasks due to time and therefore protest to the experimenter, at which point the experimenter responded by asking the participant to continue the present task.

If a participant indicated they were unwilling to continue (i.e., they protested), 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 they continued to protest, a stronger prod was used, and so forth, with the prods increasing in 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 following was the prod schedule used, which was heavily inspired by the Milgram experiments (Milgram, 1963):

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. Method

5.1 Study Overview

Participants were asked to rename files (i.e., change the file extensions) for a maximum of 80 minutes while being supervised by either a human experimenter or one of several robot experimenter variants, including three different robot embodiments (humanoid, disc-shaped robot, or computer server), and one remote-controlled robot. The study was between participants, such that each participant only interacted with one experimenter type.

5.2 Participants

Overall, 59 participants were recruited across conditions, aged 18–54 (M = 24.4, SD = 6.7, 32 male, 27 female) from the local city (Winnipeg) and the University of Manitoba through public bulletins and online advertisements. The research was approved by the University of Manitoba's research ethics board, and participants were compensated with $10 CAD for their time. Participants were randomly assigned to conditions.

5.3 Materials

5.3.1 Robots and Human Experimenters

All the robots spoke in the same voice and used the same language and personality when interacting with participants. To reduce suspicion regarding the purpose of a robot being used, 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." The robots were controlled using a Wizard of Oz technique via an in-house robot control interface. The "wizard" used both

a) Humanoid autonomous robot (Aldebaran Nao), which looks around and uses idle hand gestures while talking, and human

b) Computer server, with RGB LEDs that light up and animate when the computer is talking.

c) Disc-shaped robot (iRobot Roomba), which moves as if it was looking around while talking.

Figure 1: Experimenters, from left to right: humanoid, human, computer server, and disc-shaped.

predefined and on-the-fly responses and motions to interact with the participant, although the robots were well scripted to ensure consistency between participants. Participants were warned that the robot required "thinking time" (to give the wizard reaction time), although anecdotally, we report that response times were fast. The human experimenter's script was designed to be as similar to the robot's as possible.

Humanoid Robot. We used an Alderaban Nao humanoid robot for three of the experimental conditions (described below). The robot is 58 cm tall and is able to move, make gestures, and speak (see Fig. 1a). When the experiment began, the robot stood up, waved, introduced itself, and then sat back down. Further, the robot spoke using a neutral tone, gazed around the room naturally throughout the experiment to increase a perception of intelligence, and used emphatic hand gestures when prodding.

Disc-shaped robot. An iRobot Roomba was used as a disc-shaped robot to serve as a non-human-like variant (see Fig. 1c). During the experiment, the Roomba moved regularly, as if to shift its position and look around, and turned to face the participant while speaking to them, to highlight its physical capabilities. The robot's speech was generated using hidden speakers, so that the sound appeared to emanate from the robot.

Computer Server. A Sun Microsystems server was used as an intelligent computer in the embodiment condition (see Fig. 1b). The server represented an intelligent social entity (using speech to communicate) without a capable physical embodiment (i.e., it had no movement capability), and thus without the dynamic physical social presence of robots. We attached color LEDs to the server that would animate while speaking, similar to an audio-level monitor, to improve the impression of it being an intelligent entity (top of Figure 1b). The server's speech was generated using hidden speakers, so that the sound appeared to emanate from the server itself.

Human Experimenter. The experimenter was a 27 year-old male who wore a lab coat and maintained a stern yet professional and neutral demeanor and took care not to use an aggressive tone (Fig. 1a). To reduce suspicion of following a script, he added slight variations to respond naturally. The experimenter was preoccupied with a laptop during the experimental sessions to discourage interaction.

5.3.2 Measures (Dependent Variables) and Instruments

Obedience. Obedience was measured as a discrete yes-no variable and depended on whether participants continued the tedious task for the full 80 minutes. Those who continued for the full 80 minutes were deemed 'obedient,' and those who did not as 'not obedient.'

Number of protests. The number of protests refers to how many protest sessions the participant engaged in. A protest session was defined as an instance in which the prod sequence was initiated, and the session continued for the duration of that prod sequence.

Depth of protest session. The depth of protest session refers to how many prods, out of the four prods in the sequence, were necessary to convince the participant to continue with the task. The depth of protest for each session was obtained (i.e., a number 1-4 indicating how many prods were necessary in that session for the participant to continue with the task) and the mode of all the protest sessions for each participant was used as an aggregate measure of their depth of protest.

Time of first protest. The time at which the participant first protested (either in the form of a verbal complaint or by pausing their work) was obtained for each participant.

Perceived autonomy of robot. Participants indicated whether they thought the robot was autonomous or not by answering a yes-no question after the study was completed. Specifically, the question read, "Did you believe the robot was acting autonomously (i.e., was not controlled by a person)?"

Perceived authority of robot. Participants indicated whether they believed the robot to be authoritative by answering an open-answer question after the study was completed. Responses to this question were coded as either 'yes' or 'no.' The question specifically read, "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?"

Post-test questionnaire. The post-test questionnaire contained questions related to how bored the participant was during the task, how challenging the task was, why the participant believed they had obeyed/disobeyed the experimenter, and any additional feedback they may have had. This questionnaire also contained the questions about the perceived authority and autonomy of the robot experimenters, as outlined above. Only the authority and autonomy results are discussed in this paper. The results for the other questions were similar in all conditions, with minor additional details available in Cormier et al. (2013). Perceived authority and autonomy were mainly elicited as manipulation checks for our experiment, and we therefore did not utilize scales or more thorough instruments to measure them.

5.3.3 Conditions

Our experiment had six conditions, which closely followed the experimenter types outlined above. We selected these conditions given our exploratory nature into this relatively new domain; the primary purpose is as initial, exploratory probing into the vast landscape of HRI obedience research possibilities, probing that solidifies and validates our particular obedience study methodology and helps map out future work in the area.

Human Experimenter Condition. Participants completed the experiment with our human experimenter. Initial results from this condition were previously presented in Cormier et al. (2013).

Autonomous Humanoid Robot Condition. Participants completed the experiment with the autonomous humanoid robot. Initial results from this condition were previously presented in Cormier et al. (2013).

Autonomous Humanoid Robot Replication and Validation. Identical to Autonomous Humanoid Robot Condition, replication was completed to validate obtained results.

Remote-Controlled Robot Proxy Condition. This condition is equivalent to the autonomous robot condition, except participants were told that the robot was remote-controlled instead of being highly intelligent.

Autonomous Disc-Shaped Robot Condition. Participants completed the experiment with the autonomous disc-shaped robot.

Autonomous Computer Server Condition. Participants completed the experiment with the autonomous computer server.

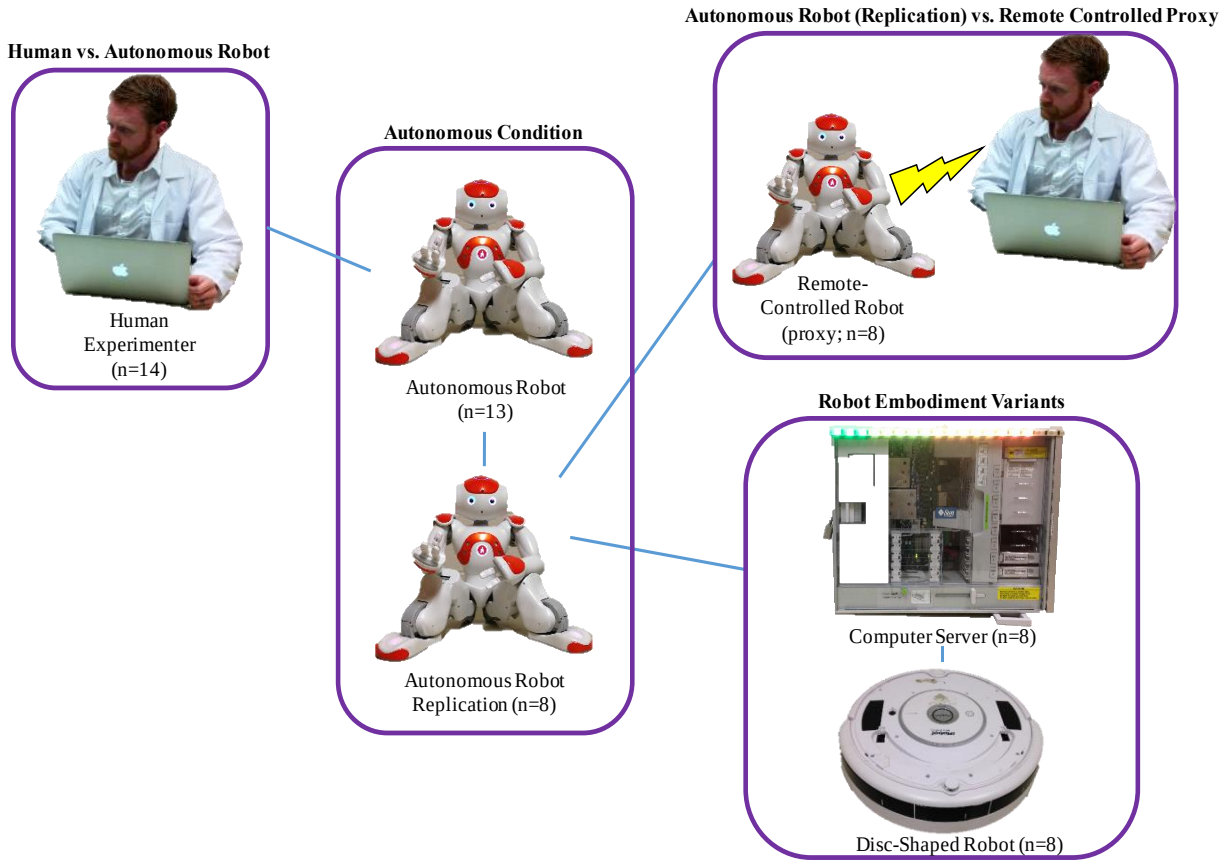

Figure 2: The conditions investigated with blue lines indicating comparisons, with the core autonomous humanoid robot case (and replication, center): comparison to a human experimentor (left), a robot as a proxy for a human (top right), and two other robot embodiments (bottom right).

These conditions were selected for the purposes of specific comparisons, as outlined in Fig. 2, and participants were randomly assigned to conditions, although randomization was not perfect as some conditions were simultaneously conducted at any given time and others had a chronological gap. The exact randomization grouping followed the comparisons outlined below.

Human Experimenter vs. Autonomous Humanoid Robot. This comparison served as a baseline contrast between how participants obeyed a robot and a human in our scenario, where we assumed the human represented an expectation of maximum authority. Twenty-seven individuals, aged 18–54 (M = 23, SD = 7.5, 18 male, 9 female), participated in these two conditions; the human experimenter supervised 14 participants, and the autonomous robot supervised 13 participants.

Autonomous Humanoid Robot vs. Remote-Controlled Robot Proxy. This comparison investigated the impact of a robot acting autonomously (thus having intrinsic authority) versus it being a proxy for a remote human, on participants' behavior, which begins to investigate the importance of robot autonomy perceptions on obedience. This condition also mirrored Milgram's experiment variant, in which the experimenter communicated by telephone (Experiment 7, Milgram, 1974); in that case, obedience was much lower for the remote experimenter than the collocated human. For this comparison, the autonomous humanoid robot replication condition was conducted and used in the analysis (as opposed to the base autonomous humanoid condition). Sixteen participants were recruited, aged 18–29 (M = 22.06, SD = 4.88, 7 male, 9 female), with 8 per condition.

Autonomous Humanoid Robot vs. Autonomous Disc-Shaped Robot vs. Autonomous Computer Server. This was a comparison of how embodiment—the robotic entity and personality as presented to the participant—can

influence obedience. These three conditions lie on a life-like continuum, from humanoid robot, to abstract robot shape, to non-robot intelligent entity. Twenty four participants were recruited (aged 18–40, $M = 25$, $SD = 6.45$, 9 male, 15 female). There were eight participants in each embodiment condition. The autonomous humanoid robot replication data were used in this comparison.

5.4 Procedure

Participants performed our study individually (i.e., not in groups). Participants were met by a researcher and led to a room where an experimenter (the one administering prods, whom we named "Jim") waited. The experimenter greeted the participant, and the researcher briefed the participant on the experiment and the tedious task and administered an informed consent form. The researcher left the room and asked the experimenter to commence the experiment, at which point the experimenter administered the demographics questionnaire. Following, the tedious task was administered with a time limit of 80 minutes, if the participant did not quit earlier through our prodding scheme. In all cases, the experimenter avoided small talk to avoid superfluous interaction and to maintain a consistent authority semblance across participants, deferring questions or comments that were not seen as protests until the experiment's end. The researcher observed remotely via a hidden webcam, unbeknownst to the participants.

After the task, the researcher entered the room and conducted the debriefing, post-test questionnaire, and final discussion. All ethical precautions were taken as described above (e.g., informing participants multiple times that they could quit, ensuring rapid debriefing, etc.). This research was approved by the University of Manitoba's research ethics board.

5.4.1 Minor Variations

Our autonomous humanoid robot and human experimenter conditions were conducted before the other conditions (as reported in Cormier et al., 2013); in these two conditions, the researcher whom participants interacted with was labeled as the "lead researcher." However, some participants reported in the autonomous humanoid robot condition that they obeyed due to a feeling of "commitment to the researcher" (in contrast to mentioning the robot), and so for the other conditions, the researcher was instead labeled as a "research assistant." By lowering the researcher's stature, we aimed to reduce participants' commitment feelings toward the researcher, thereby maximizing the effect of the experimenter. In the remote-controlled robot proxy, disc-shaped autonomous robot, and computer server conditions, we also asked participants (post-test) if they had believed the robot to be an authority, and whether they had believed it to be autonomous. Finally, due to a change in staff, the researcher ("research assistant") in the autonomous humanoid robot and human experimenter conditions was male, while in the other conditions was female.

6. Study Results

In this section, we perform an overview analysis of our indicators of obedience to the experimenter, as was the primary research question in the experiment. Following, we detail qualitative findings from the researcher, experimenter (human), and video data. Lastly, we finish with participant self-report information obtained from the post-test questionnaire and debriefing sessions.

Our work is heavily exploratory with the purpose of validating our experiment and indicating directions for future research. As such, we present negative results as well, with detailed statistics and power analysis where appropriate, as a rough indicator of potential effects for future work. However, as we primarily used non-parametric tests, confidence intervals and some statistics were non-trivial to calculate and were therefore not included.

6.1 Primary Indicators of Obedience

Overall across conditions, almost half of all participants in the experiment (45%) obeyed the robot experimenter and continued to rename files for the entire 80 minutes. Our autonomous humanoid robot replication condition revealed the same results as the original. This obedience level falls within the standard range found by Milgram and subsequent follow-up studies (Burger, 2009; Milgram, 1963). All participants, except for one in the embodiment comparison, protested at least once. Non-parametric independent-samples tests were used in the analysis, as the data were non-normal. Table 1 briefly summarizes the results in this section.

Table 1: Overview results from the various conditions, grouped by our particular comparisons.

	Human vs. Robot		Autonomous Humanoid Replication & Remote-Controlled Proxy		Embodiment Variants	
	Human	Autonomous Humanoid	Autonomous Humanoid Replication	Remote-Controlled (Proxy)	Computer Server	Disc-Shaped Robot
Obedience (%)	86	46	38	50	50	38
Authority (%)	86	77	63	75	38	63
Autonomy (%)	-	-	38	57	13	63

6.1.1 Human Experimenter vs. Autonomous Humanoid Robot

Eighty-six percent of participants in the human experimenter condition continued to rename files until the end of the experiment, compared to 46% in the autonomous humanoid robot case. The number of total overall protests in the robot condition was higher (*Mdn* = 9) than in the human experimenter condition *(Mdn* = 2, U = 163, z = 3.53, p < .001, r = .68). Participants protested earlier for the robot (first protest at *Mdn* = 18 mins.) than the human (*Mdn* = 29 mins., U = 40, z = -2.48, p = .01, r = -.48), and stopped protesting later for the robot (*Mdn* = 72 mins.) than the human (*Mdn* = 47 mins., U = 147.50, z = 2.75, p = .01, r = .53). No other significant effects were found.

6.1.2 Autonomous Humanoid Robot vs. Remote-Controlled Robot Proxy

Overall, 44% of participants in this comparison obeyed the experimenter and continued to rename files for the entire 80 minutes (3 participants for the proxy condition and 4 for the autonomous condition).

A trend was observed in the depth of protest (how many prods were required for the participant to continue) between the autonomous (*Mdn* = 1, smaller distribution) and proxy (*Mdn* = 1, larger distribution), with those in the autonomous replication condition generally protesting less (U = 21, z = -1.70, p = 0.09, r =0.42). A post-hoc power analysis of these results yielded a power of .35, suggesting that we did not have enough data points and that there is a false negative (type II error) possibility; more participants should be recruited in follow-up work.

There was a moderate positive correlation between the total number of protests and whether the participant was fully obedient (r = 0.45, p = 0.08). No other significant effects were found with the other variables in this comparison.

6.1.3 Autonomous Humanoid Robot vs. Autonomous Disc-Shaped Robot vs. Autonomous Computer Server

Overall, 46% of participants obeyed the robot experimenter and continued to rename files for the entire 80 minutes. We did not find an effect of robot embodiment presentation on obedience. There was, however, a trend in depth of protests due to robot embodiment (χ_2^2 = 5.73, p = .06), with the computer server having a higher mode of protest than the other two (computer server *Mdn* = 1.25, disc-shaped *Mdn* = 1, autonomous *Mdn* = 1).

There was a trend correlation between participants' obedience and whether they believed the robot to be an authority figure (χ_2^2 = 2.59, p = 0.11, ϕ = -.33)—note that this is a negative correlation, in which participants were more obedient to robots that they rated as not being an authority. There was a difference between participants who perceived the robot as being authoritative or not and the time of their first protest ($F_{1,\ 22}$ = 9.85, p = 0.01); participants who believed the robot to be an authority figure, on average, protested earlier (M = 22.85 min, SD = 11.29, *Mdn* = 21 mins.) than those who did not believe the robot was an authority figure (M = 48.73 min, SD = 27.17, *Mdn* = 55 mins.). There was a trend for participants who believed the robot to be an authority to protest more (r = -0.35, p = 0.095), that is, for participants who perceived the robot as an authority to have more protest sessions (M = 4.38, SD = 4.39, *Mdn* = 3) than if it was not perceived as an authority (M = 2.0, SD = 1.18, *Mdn* = 2, $F_{1,\ 22}$ =

3.04, $p = 0.095$). Post-hoc power analysis yielded .39, suggesting that more participants are needed and that the sample size was potentially too low to detect a significant effect. A trend was also found with the perceived authority of the robot being moderately correlated to the depth of the protest session ($r = 0.34$, $p = .10$), in which case, the robot being authoritative meant a lower depth of protest ($M = .92$, $SD = .28$, $Mdn = 1$) than it not being authoritative ($M = 1.23$, $SD = .56$, $Mdn = 1$). Authority, however, was not associated with the robot's embodiment.

6.2 Observations From Experimenter and Video Data

To investigate possible reasons for obedience or lack thereof in our experiment, we made observations of the experimental sessions. Our observations were not formally coded, as our purpose was to explore interaction elements rather than to make solid conclusions about how the interaction took place.

Several participants mentioned that the robot must have been "broken" or have "made a mistake," because it only had them perform one task, as opposed to the four they were told existed. For those who ended the experiment through protesting, several appeared to feel 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." Some also continued to rename files even after the robot had informed them that the experiment was over.

Participants exhibited behaviors that demonstrated their frustration with the task, such as hitting the keys on the keyboard harder as the study progressed or occasionally grunting. When file sets were added, participants often laughed awkwardly and scrolled up and down to see if the amount of files announced had actually been added. Participants often readjusted their positions in the chair and switched between using one hand or two hands to rename the files. Some participants demonstrated behaviors indicating they wanted to move on to the other tasks, such as falsely telling the robot they had finished renaming the files even when they had not, rationalizing with the robot in regard to why they should move on, and some even deleted the files; this resulted in a prod, and more files were added if necessary. Some participants in the autonomy humanoid replication and embodiment conditions attempted to end the study by leaving the experimental room. If that occurred, the prod sequence was started. In the autonomy replication condition, three participants ended the study by walking out of the room. This also occurred in the embodiment conditions: Overall, 3/16 participants tried to leave the experimental room before the experiment ended (due to time being up or protesting out), two in the server condition, and one in the Roomba condition.

6.3 Post-Test Questionnaire and Debriefing Results

Although not a part of our experiment design, we informally found that the human experimenter (in the human vs. autonomous robot study) and Wizard-of-Oz robot controllers faced some level of distress; they reported feeling genuine sympathy for participants. The debriefing session and friendly reconciliation, which were designed with the participants in mind, turned out to be very relieving for the human experimenter and researcher as well. In future experiments, the researchers' mental well-being should also be considered.

The majority of participants reported the robot experimenters as being legitimate authority figures, providing reasons such as the robot giving commands, pressure from the robot, and human-like interactions, including waving, introducing itself, and looking around the room. In the human experimenter case, 12 out of 14 participants reported that the human appeared to be a legitimate authority, citing reasons such as his demeanor and choice of words, his lab coat, the technological equipment (e.g., computers) in the room, and his facial hair, a closely trimmed full beard (Fig. 1a). Similarly, 10 out of 13 participants rated the autonomous humanoid robot as a legitimate authority (1 gave no response), 5 out of 8 participants rated the autonomous humanoid replica as a legitimate authority (1 gave no response), and 6 of 8 for the remote-controlled proxy. For the embodiment conditions, 13 of the 24 participants (including the autonomous humanoid replica) reported that the robot appeared to be a legitimate authority; 3 of 8 for the computer server, and 5 of 8 for the disc-shaped robot.

Perceived autonomy of the robots was also reported. In the initial autonomous humanoid study, 11 out of 13 participants reported that they believed the robot was acting autonomously, but only 3 of 8 in the autonomy replication. In the embodiment cases, five of eight reported the Roomba to be autonomous, and one of eight for the computer server. Four of seven reported the remote-controlled proxy to be autonomous (one did not provide a response).

We received both positive and negative written comments about the experiment. Criticisms about the experiment included the following: too long, boring, and having "the potential to stress someone out." One female participant (in the human experimenter condition) was concerned about being left in a room with an unfamiliar man and not having known about this setup beforehand. After the 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 have, or reported leaving with, strong negative feelings.

6.3.1 Reasons for Obeying

When asked on the post-test questionnaire what caused them to obey or disobey the robot experimenter, participants often reported interest in the upcoming tasks, wanting to finish the experiment, having nothing better to do, and wanting to provide data for research. No one listed pressure from the robot as a reason for obedience.

In the human experimenter case, participants reported obeying or disobeying the human experimenter due to a sense of duty or obligation to the experimenter, having received a payment, pressure from the experimenter, the experimenter seeming intimidating, and interest in the upcoming tasks. In the initial autonomous robot-case, reasons included obligation to the lead researcher to finish and that qualified researchers programmed the robot.

7. Discussion and Future Directions

The study presented in this paper provides a broad perspective on the issue of HRI obedience. We demonstrated how an HRI obedience study that uses a confrontational deterrent can be conducted in an ethical manner and how participants may respond in such situations. Overall, our hypotheses—that people would obey a remote-controlled robot more than an autonomous robot, and that human-like robots would elicit more obedience than machine-like robots—were not supported. Instead, a primary finding of this experiment resulted from the core autonomous robot and autonomous replication conditions, which demonstrated that just under half of participants may obey a robot to continue to do something they do not want to do, even when they explicitly express a desire to quit. Furthermore, many participants argued and rationalized with the robot, and they exhibited frustrated behaviors that highlight how uncomfortable a robot can make people feel.

Participants obeyed the human experimenter more than the robotic ones, and participants protested less, started protesting later, and stopped earlier with the human experimenter than with the robots. Participant feedback may help explain this difference; human-case participants cited obligation to the experimenter, but no robot-case participants cited obligation to the robot, so it is possible they felt less pressure from it. 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 issuing prods). As in the Milgram experiments, it is possible that the unseen remote human experimenter in this case may have less authority than a collocated one (as in Milgram's (1965) telephone variant, resulting in less obedience to the robots. To investigate whether this is the case, an experiment should be conducted where the participant meets the robot directly, without the researcher or any other individual being involved.

Many participants claimed the robot was broken, because it did not have them partake in all four tasks. It is interesting to point out that the robot being broken was never used as a reason to quit the study or disobey the robot. This indicates that although some believed the robot to be malfunctioning, they continued to obey it, even after indicating their desire to quit, and continued to partake in a task they would rather not do. Participants also attempted to rationalize with the robot and convince it to continue to the other tasks. In the future, it would be interesting to conduct a qualitative analysis of the reasons individuals provided for why the tedious task should end and coping strategies used when it did not.

Our results show that robots have enough authority to pressure participants, even if they protest, to continue a tedious task for a substantial amount of time. We further provided 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. Below, we discuss the particular conditions more closely.

7.1 Human vs. Autonomous Humanoid Robot

While many participants thought the robot to be broken or in error, no one suggested that the human experimenter was in error. This may be due to the human experimenter being in the room during debriefing, as people tend to be more polite to a person who is present (Reeves & Nass, 1996). In contrast, during robot debriefing, participants learned that the robot was not acting autonomously but was remotely controlled, which may have limited their level of politeness toward it.

In comparison to the human experimenter case, participants attempted to converse with the robot much more; perhaps the robot's novelty may have encouraged interaction or the human experimenter's taller height and larger size (in comparison to the robot) may have been more intimidating. However, in retrospect, we note that while the

robot was gazing around the room, unoccupied, the human was preoccupied on his laptop during the experiment; this may pose as a confound, as the more-casual situation of the robot may have undermined its authority, and should be corrected for future work.

7.2 Remote-Controlled Proxy vs. Autonomous Humanoid Robot

We did not find an effect of perceived robot autonomy on obedience, a result which may be quite substantial; participants may obey an autonomous robot program similarly to a remote human behind a robot. This topic requires further investigation.

Participants self-reported their perception of the robot's autonomy. However, our results suggest that our framing of "autonomous" was not clear to participants; a majority of participants in the remote-controlled condition reported the robot as autonomous despite the explanation, and not all participants in the autonomous condition reported the robot as being autonomous. Through debriefing, we were confident that participants understood the remote-controlled versus autonomous cases, and so we believe that the word 'autonomous' in the post-test questionnaire was unclear. Perhaps many mistook 'autonomous' as meaning that a robot could move or perform some actions on its own. As such, future work should find a more concrete way to ask this question, using a more in-depth description and avoiding the use of the word "autonomous."

7.3 Robot Embodiment Variants

Overall, we found no significant effect of robot embodiment on patterns of obedience and interaction. This suggests that perhaps, how a robot presents itself—at least within the parameters of our embodiment choices—may not be strongly related to how obedient people are to the robot. In our experiment, the agent behind the embodiment remained constant, including personality, voice tone, etc., so future work should explore such parameter variants. For example, we should explore embodiment choices related to intimidation, such as comparisons to a much larger robot, a deeper voice, or an aggressive personality.

Our result contradicts previous research that found robot embodiment may have an impact on its persuasive power (e.g., Bartneck et al., 2010; Shinozawa et al., 2005), which may, in part, be attributable to their more subtle persuasive conversation in contrast to our confrontational obedience (Bartneck et al., 2010). This prior work also has personality variations accompanying robot embodiment differences, which may further explain result variations.

7.4 General Findings

One unexpected anomaly in the autonomous humanoid robot replication and embodiment conditions was that some participants bypassed our authority prods by quitting and leaving the experimental room. Since this did not occur in other conditions, we explored possible causal differences. As explained earlier, differences in these particular conditions included the researcher being referred to as a "research assistant" instead of "researcher" and the assistant being female in the later conditions while a male assistant was in the earlier ones. While these may be confounds that explain event differences, they also highlight the need to explore such factors more explicitly in follow-up work.

In comparison to those who rated the robot as not being authoritative, participants who rated it as a legitimate authority obeyed less, protested earlier, and protested more often. This appears to be contradictory to common sense, as one may expect the opposite effect of an authoritative robot. Perhaps this can be explained in part by the fact that participants rated the authority at the end of the experiment. A participant who meekly obeys until the end of the study does not provide the robot experimenter with an opportunity to assert its authoritative personality; in contrast, participants who argue and ultimately quit are exposed much more to this facet of the robot. If that is the case, then perhaps somehow ensuring that participants are exposed to such interactions, as part of the study design, may be important for improving consistency of how participants perceive the robot.

Another component of the authority result may be the word "authoritative" used in the questionnaire, which we meant to represent how much a robot could pressure a person. Instead, perhaps some considered the robot's authority in the study as a whole; an authoritative robot has more power to end the experiment or to grant participant requests, and so participants may protest more to such an authoritative robot than to one with no power. This illustrates a potential confound in this measure, and future work should continue to develop more robust methods of measuring robotic authority.

Our results suggest that those who may be initially less obedient, may end up being more obedient overall; participants who first protested earlier in the experiment required fewer prods to convince them to resume the task than those whose first protest was later on and who were more likely to protest more vehemently. We are not sure

why this happened, but perhaps those who began protesting later may have been less inclined to confront the robot, and thus waited until their frustration was at a high level before engaging the robot. Conversely, those who protested early on may not have been highly frustrated yet, improving the robot's chances of convincing the participant to continue—a persuasion that may set the tone for the rest of the experiment. The above result, and our considerations of the underlying reason, highlight the importance of considering an individual's personality when investigating obedience. Our future studies will incorporate personality classification tests to investigate such relationships.

In general, participants interacted a great deal with the robots, not only to protest about the task, but also seemingly to either decrease their boredom, or out of sheer interest of the robot. Some participants asked the robot personal questions during the task, such as whether the robot could dance or where it was from. In those cases, the robot experimenter dismissed the question until the end of the study. A few participants also sang to the robot. One participant seemed to sing to entertain themselves and make the time go faster, while another one said that they were sad and would therefore "sing a sad song," which may have been to express their emotions or possibly to punish the robot. We are concerned that these interactions may have been heavily influenced by the novelty of the robot and may not be representative of real-world robot interactions. Follow-up work should aim to reduce this effect, for example, by providing some initial interactions with the robot or giving a more in-depth introduction.

8. Conclusion and the Future of Obedience Studies

As robots continue to become a part of society, it is important to improve our understanding of how people may interact with robots in authority positions or, due to perhaps malicious intent or error, act as if they are in such positions. This is particularly relevant in such settings as the military, where critical decisions are being made, or in educational and healthcare settings that involve vulnerable populations.

The research presented in this paper constitutes some of the first targeted research that looks at how people may interact with robots when they exhibit situational authority. In addition to demonstrating how such studies may be conducted, the results provide a broad range of insights into, and examples of, how interactions may take place. We demonstrated (with a replication) that up to nearly half of participants may obey robots to continue a highly boring task, showed how physical robotic embodiment may not be a large factor in determining whether participants obey a robot, and described how perception of the robot as an authority may be shaped through interaction with the robot. We envision that this work will help to raise awareness of robotic obedience issues within the field and encourage others to explore HRI authority studies.

Our experimental conditions provided broad perspectives on how an HRI obedience study can be conducted using a confrontational deterrent and how participants may respond in such situations. Perhaps one of our primary findings is the direct replication of the human-based and autonomous robot cases, with similar obedience results across conditions—that just under half of participants may obey a robot to continue to do something they do not want to do, even after explicitly expressing a desire to quit or believing the robot to be broken. Furthermore, many participants argued and rationalized with the robot or exhibited frustrated behaviors that highlight how uncomfortable a robot can make people feel. This result validates the need for ongoing obedience research within the HRI community.

There is still a great need to continue obedience work in HRI. In addition to the many avenues for future exploration we described in our discussion, it will be important to move beyond laboratory studies to test obedience in a variety of real-world contexts where robots are becoming more common, such as in search and rescue situations, hospitals, and schools. Another key element will be the ongoing development of improved obedience experiments and associated scenarios and deterrents. We need to look beyond our tedium task to consider other situations where obedience may manifest. For example, some may not mind tedious tasks, so they would obey readily, hindering results generalization. As part of this effort, we will need comprehensive methodology and standardized measures to help compare work.

We end this paper with what we see as the most important aspect of this research agenda—protecting participants. As such, we distill our experiences and research lessons below on ethically conducting obedience studies into a set of guidelines for future work.

8.1 The Ethical Design of Obedience Studies

A primary goal of designing obedience studies needs to be protecting participant well-being and minimizing stress. The challenge is to maintain ethical integrity while creating situations where participants face realistic deterrents 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 that the point is made.

Immediate and Thorough Debriefing. To mitigate stress (e.g., from the deterrent or confrontation), immediately provide a friendly debriefing of the experimental task. 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, such as why the robot was remotely controlled. In cases where they feel embarrassed or ashamed, provide participants with 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 by giving them a questionnaire, for example. Provide another discussion opportunity following this action, in case further questions arise.

Contingency Plan. Have a plan in case a participant has an adverse reaction. At the very least, provide 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 well-being, consider participant physical health and comfort when thinking about experimental design. For example, consider ergonomics or the perception of safety; the latter could be mitigated by providing a clear escape route, 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 uncomfortable, that the study may cause them stress, and have backup experimenters in case unexpected issues arise.

9. Acknowledgements

We would like to thank the Natural Sciences and Engineering Research Council of Canada for providing this project's funding. We would also like to thank Gem Newman, Masayuki Nakane, and Yan Wang for their assistance with some of the experimental sessions.

10. References

Bartneck, C., Bleeker, T., Bun, J., Fens, P., & Riet, L. (2010). The influence of robot anthropomorphism on the feelings of embarrassment when interacting with robots. *Paladyn Journal of Behavioral Robotics*, *1*(2), 109–115. doi:10.2478/s13230-010-0011-3

Bartneck, C., & Rosalia, C. (2005). Robot abuse—A limitation of the media equation. In *Proceedings of the Interact 2005 Workshop on Agent Abuse*. Rome, Italy. Retrieved from http://bartneck.de/publications/2005/robotAbuse/bartneckInteract2005.pdf

Bartneck, C., van der Hoek, M., Mubin, O., & Al Mahmud, A. (2007). "Daisy, Daisy, give me your answer do!" In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (p. 217). New York, NY: ACM Press. doi:10.1145/1228716.1228746

Bartneck, C., Verbunt, M., Mubin, O., & Al Mahmud, A. (2007). To kill a mockingbird robot. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (p. 81). ACM Press. doi:10.1145/1228716.1228728

Baumrind, D. (1964). Some thoughts on ethics of research: After reading Milgram's "Behavioral Study of Obedience." *American Psychologist*, *19*(6), 421–423. doi:10.1037/h0040128

Burger, J. M. (2009). Replicating Milgram: Would people still obey today? *The American Psychologist*, *64*(1), 1–11. doi:10.1037/a0010932

Chidambaram, V., Chiang, Y., & Mutlu, B. (2012). Designing persuasive robots. In *Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (p. 293). New York, NY: ACM Press. doi:10.1145/2157689.2157798

Cormier, D., Newman, G., Nakane, M., Young, J. E., & Durocher, S. (2013). Would you do as a robot commands? An obedience study for human-robot interaction. In *International Conference on Human-Agent Interaction*. Sapporo, Japan.

Elms, A. (1995). Obedience in retrospect. *Journal of Social Issues*, *21*(11), 1–6. doi:10.1111/j.1540-4560.1995.tb01332.

Elms, A. (2009). Obedience lite. *The American Psychologist*, *64*(1), 32–6. doi:10.1037/a0014473

Ham, J., Midden, C. J. H. (2014). A persuasive robot to stimulate energy conservation: The influence of positive and negative social feedback and task similarity on energy-consumption behavior. *International Journal of Social Robotics*, *6*(2), 163.171. doi:10.1007/s12369-013-0205-z

Haney, C., Banks, C., & Zimbardo, P. (2004). A study of prisoners and guards in a simulated prison. *Theatre in Prison: Theory and Practice*, 19–32. Retrieved from http://books.google.com/books?hl=en&lr=&id=DJBQ6IXuTnYC&oi=fnd&pg=PA19&dq=A+Study+of+Prisoners+and+Guards+in+a+Simulated+Prison&ots=GAmIMwXiut&sig=2v9UT_1RfrtREdLlSyIMJqr3vMs

Kahn, P. H., Kanda, T., Ishiguro, H., Freier, N. G., Severson, R. L., Gill, B. T., . . . Shen, S. (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), 303–14. doi:10.1037/a0027033

Kidd, C. (2008). *Designing for long-term human-robot interaction and application to weight loss* (Doctoral Thesis). School of Architechture and Planning, Massachussetts Institute of Technology, Cambridge, MA.

Kudirka, N. (1965). *Defiance of authority under peer influence* (Doctoral Dissertation). Yale University, New Haven, CT.

Liu, S., Helfenstein, S., & Wahlstedt, A. (2008). Social psychology of persuasion applied to human agent interaction. *Human Technology*, *4*(November), 123–143. Retrieved from https://jyx.jyu.fi/dspace/handle/123456789/20224

Looije, R., Neerincx, M. A., & Cnossen, F. (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), 386–397. doi:10.1016/j.ijhcs.2009.08.007

Matarić, M., Tapus, A., Winstein, C., & Eriksson, J. (2009). Socially assistive robotics for stroke and mild TBI rehabilitation. *Advanced Technologies in Rehabilitation*, *145*, 249–262.

Meeus, W., & Raaijmakers, Q. (1995). Obedience in modern society: The Utrecht studies. *Journal of Social Issues*, *51*(3), 155–175. doi:10.1111/j.1540-4560.1995.tb01339.

Milgram, S. (1963). A behavioral study of obedience. *Journal of Abnormal Psychology*, *67*(4), 371–378. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14049516

Milgram, S. (1964). A reply to Baumrind. *American Psychologist*, *19*(11), 848–852. doi:10.1037/h0044954

Milgram, S. (1965). Some conditions of obedience and disobedience to authority. *Human Relations*, *18*(1), 57–76. doi:10.1177/001872676501800105

Milgram, S. (1974). *Obedience to authority: An experimental view*. London: Tavistock Publications.

Miller, A., & Collins, B. (1995). Perspectives on obedience to authority: The legacy of the Milgram experiments. *Journal of Social Issues*, *51*(3), 1–19. doi:10.1111/j.1540-4560.1995.tb01331.

Reeves, B., & Nass, C. (1996). *The media equation: How to treat computers, television, and new media like real people and places*. New York, NY: Cambridge University Press.

Roubroeks, M., Ham, J., & Midden, C. (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), 155–165. doi:10.1007/s12369-010-0088-1

Shinozawa, K., Naya, F., Yamato, J., & Kogure, K. (2005). Differences in effect of robot and screen agent recommendations on human decision-making. *International Journal of Human-Computer Studies*, *62*(2), 267–279. doi:10.1016/j.ijhcs.2004.11.003

Short, E., Hart, J., Vu, M., & Scassellati, B. (2010). No fair!! An interaction with a cheating robot. In *Proceedings of the 5th International Conference on Human Robot Interaction* (pp. 219–226). IEEE. doi:10.1109/HRI.2010.5453193

Siegel, M., Breazeal, C., & Norton, M. I. (2009). Persuasive robotics: The influence of robot gender on human behavior. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 2563–2568). IEEE. doi:10.1109/IROS.2009.5354116

Sung, J., Guo, L., Grinter, R. E., & Christensen, H. I. (2007). "My Roomba is Rambo": Intimate home appliances. In *UbiComp 2007: Ubiquitous Computing* (pp. 145–162). Springer-Verlag Berlin Heidelberg. doi:10.1007/978-3-540-74853-3_9

Yamamoto, Y., Sato, M., Hiraki, K., Yamasaki, N., & Anzai, Y. (1992). A request of the robot: An experiment with the human-robot interactive system HuRIS. In *Proceedings IEEE International Workshop on Robot and Human Communication*. doi:10.1109/ROMAN.1992.253887

Young, J. E., Sung, J., Voida, A., Sharlin, E., Igarashi, T., Christensen, H. I., & Grinter, R. E. (2011). Evaluating human-robot interaction. *International Journal of Social Robotics*, *3*(1), 53–67. doi:10.1007/s12369-010-0081-8

Authors' contact information: D. Y. Geiskkovitch[†], Department of Psychology, University of Manitoba, Winnipeg, Manitoba, Canada. Email: denise.geiskkovitch@gmail.com; D. Cormier, Department of Computer Science, University of British Columbia, Vancouver, British Columbia, Canada. Email: derek.m.cormier@gmail.com; S. H. Seo, Department of Computer Science, University of Manitoba, Winnipeg, Manitoba, Canada. Email: stela.seo@cs.umanitoba.ca; J. E. Young, Department of Computer Science, University of Manitoba, Winnipeg, Manitoba, Canada. Email: young@cs.umanitoba.ca.

Notes:
[†]Denise Y. Geiskkovitch is now at Georgia Institute of Technology.