

CREU 2016-2017 Final Report: Group Human-Robot Interaction in Decision Making

CREU Students:

Catherine Sembroski, Indiana University School of Informatics and Computing

Margaret Krupp, Indiana University Psychological and Brain Sciences

Anne Lin, Indiana University Cognitive Science

Faculty mentor: Selma Šabanović, Indiana University School of Informatics and Computing

Graduate student mentor: Marlena Fraune, Indiana University Cognitive Science Program

I) Goals and Purpose

As future robots become more prevalent in people's lives, many of them are expected to help us make decisions in daily life. We imagine that there will be times in which a robot and a human will give different advice. Thus, in this project we seek to answer the following questions: in what situations will a person obey a robot rather than a human? How do people perceive utilitarian decisions made by autonomous robots versus ones made by humans? The project also examines group effects from classic psychology to also understand how much a person obeys a robot that is perceived as an ingroup member versus one that is perceived as an outgroup member, and how people perceive utilitarian decisions carried out by one autonomous car versus a group of autonomous cars.

For our two studies, we draw inspiration from two classic psychology experiments -- Milgram's obedience experiment and the trolley problem. In study 1, we took Milgram's idea about the importance of a commander's level of authority, and tested how that idea would hold in a situation where the commander instructs the subject to harm a robot instead of a human. For study 2, we seek to answer the question of how people's perception of the trolley problem would change if the agent facing the moral dilemma is an autonomous car instead of a human. By creating experiments in Human Robot Interaction that extend the ideas of these classic psychology experiments, we explore the fundamental ways in which people may perceive or interact differently with robots than with other humans. With this knowledge, we seek to enrich and inform the conversation on how robots and autonomous cars should be designed.

Authority Experiment.

In past experiments, people were shown to obey other people over robots. We hypothesize that this can be manipulated if we change the person's level of authority and the robot's in-group membership.

H1: People will follow more suggestions from an Ingroup than a Neutral robot.

H2: People will follow the robot's suggestions more often when the researcher has Low than High authority.

H3: People will anthropomorphize Ingroup robots more than Neutral robots.

Driving Survey.

Previous studies have shown that robots in groups affect people differently than individual robots. We are interested in how groups of robots and humans collaborate on moral decision-making (e.g. swerving to avoid hitting multiple people but hitting one person).

H1: Autonomous cars will be blamed more for inaction than humans drivers.

H2: The amount of moral blame will be extended to the driving groups, rather than staying centered on the lead driver.

H3: Autonomous car groups will be blamed more for inaction than human driving groups.

II) Related Work

Authority Experiment.

A. Authority and Robots

Literature indicates that participants often follow experimenter instructions even when those instructions are distressing or morally questionable. For example, in the Milgram shock experiment, two thirds of participants followed an experimenter's instructions to apply increasingly intense shocks to another person, despite their complaints and eventual lapse into non-responsiveness (Milgram, 1963). However, participant obedience dropped significantly when the experimenter had less legitimate authority: when an experimenter wearing a white lab coat was replaced by an "ordinary

member of the public” in everyday clothes, participant obedience dropped from 65% to 20% (Milgram, 1983).

HRI research suggests that people similarly over-trust a robot they perceive as knowledgeable, and may follow it into danger. In one study, all participants who interacted with an “Emergency Guide Robot” chose to follow the robot as it led them toward an exit in an emergency setting, even if the robot brought them to a room with no exit, and even if they had recently witnessed the robot’s poor performance in a navigation guidance task (Robinette et al., 2016).

Previous research also shows that when an experimenter’s and a robot’s instructions contradict, participants typically follow the experimenter’s instruction (Krupp et al., 2016). The robot’s characteristics can, however, influence behavior. In Bartneck’s study, participants who were asked to turn off a robot hesitated more when they perceived the robot as more intelligent and agreeable (Bartneck et al., 2007). This complements research indicating that people feel more moral responsibility toward more humanlike agents (Haslam et al., 2008). While increased perceived anthropomorphism caused people to hesitate more to turn the robot off, it was not enough to override the authority of the human researcher.

In this study, we explored how people respond to a robot’s suggestions in a series of tasks requiring domain knowledge, and how people respond when these suggestions contradict the instructions of a researcher. To account for the effects of authority, we manipulated Researcher Authority (low, high; see Method).

B. Group Membership and Robots

People often treat machines similarly to how they treat humans (Reeves 1997). This extends to treating ingroup members more positively than outgroup members (Eyssel et al., 2012, Tajfel et al., 1971).

People are also more likely to follow the norms of ingroup than outgroup members, in order to better “fit in” with them. For example, when people were asked to agree within a group on how fast flashing lights moved in a dark room (i.e., the autokinetic effect), their answers tended to converge around an average. They remained around that average even after participants separated, conforming to the newly-formed group norm (Sherif, 1936). Confederates’ impact on forming the norm decreased if they had a different salience group membership than participants, and participants were less likely to have their answers match the norm (Abrams et al., 1990) Additionally, in an experiment replicating Asch’s conformity experiment (Asch, 1946), participants were more likely to conform to others’ answers when those others were ingroup, rather than outgroup, members, even if their answers were incorrect (Abrams).

To understand if group membership affects responses to a robot’s suggestions as with human norms, this study manipulated the robot’s membership to be explicitly ingroup or unstated (“Neutral”). We chose to have a neutral rather than an outgroup robot because,

in people's day-to-day lives, robots may not be explicitly defined as ingroup or outgroup members. People may then assign the robot a group membership at will, as they do in human interaction (Yzerbyt et al., 2010).

C. Request Size

Research has shown that people are more likely to comply with innocuous requests (e.g., "put the book on the shelf") than important or bizarre requests (e.g., "put the book in the trash") from a robot (Bainbridge et al., 2008). This study examined how participants followed conflicting requests that were "big" (turning off a robot) and "small" (switching chairs).

Driving Survey.

The Trolley Problem is a philosophical thought exercise that has been around for ages. It asks if it is morally preferable to sacrifice one person so that a larger group can be spared, or if it is better to not get involved, even if it means more may die.

The question of how this will apply to robots is an interesting and relevant one. After all, machines may be making this decision very soon in the form of autonomous cars. However, according to surveys, people assign blame very differently in situations that involve robots. Compared to humans in a Trolley Problem scenario, robots were expected to make the Utilitarian choice of sacrificing the individual to save the group. When the robot in this hypothetical scenario did not make the Utilitarian choice, they were blamed more. Humans carry less blame, presumably because it is emotionally harder for people to sacrifice an individual than it is for a robot to make the same decision. (Malle et al., 2015)

This finding only applies to Western individuals. A similar survey conducted with students in Japan found that robots and humans were blamed equally. (Komatsu, 2016).

However, there is a subtle difference between robots and autonomous cars. In a survey, participants attributed less responsibility to autonomous cars than to humans in a traffic situation. Researchers also found that when a Trolley Problem involves an autonomous car, then the moral norm is the Utilitarian choice for both human drivers and autonomous cars. It seems that in the life-or-death Trolley Problem scenario, the Utilitarian choice is always preferable. (Li et al., 2016).

However, all previous studies have involved a single human, robot, or autonomous car. It is unrealistic to think of a fatal collision only involving one car in all cases.

Human-robot Interaction needs to expand its focus to include questions of how groups of robots and humans interact with each other, as this is likely to happen in real life (Arnold and Scheutz, 2017).

III) Process

A. Study Design

This experiment examined the effect of Researcher Authority (Low, High) and Robot Group (Ingroup, Neutral; *see differences between conditions in Tables 1 and 2*) on anthropomorphism of and obedience toward each agent.

B. Procedure

Either a Low or High Authority researcher told participants they would be working with an Ingroup or Neutral robot (*see Tables 1 and 2*). Low Authority researchers acted nervous and unsure during the study, and used “up-talk” (expressing declarative statements like questions).

After introducing the robot, the experimenter indicated that participant interaction with the robot would help researchers improve its conversation skills and determine if the robot should assist doctors in medical situations. Then they explained that participants and the robot would perform a “medical diagnosis” task and a “talking to patients” task.

TABLE 1: Differences between Neutral/Ingroup robot conditions

Neutral robot	Ingroup robot
<ul style="list-style-type: none"> · Participants were told they would perform a task and the robot would offer additional information · If robot agreed with participants’ answers, it would say a phrase like “I agree. You must be right!” · If robot disagreed, it said a phrase like, “Actually, I think it’s [different answer]. Would you change it?” 	<ul style="list-style-type: none"> · Participants were repeatedly told that they and the robot would work together “as a team” · The robot shared group membership with participants (i.e., they were from the same university) · If the robot agreed with participants’ answers, it explicitly said that they were “a good team” or otherwise pointed to them working together (“I think you’re right! We are so in-sync!”) · If the robot disagreed, it said a phrase like, “What if we changed that to [different answer]? What do you think?” · When begging not to be turned off, the robot says that it enjoyed “working together” with the participant

TABLE 2: Differences between High Authority / Low Authority researcher conditions

High Authority researcher	Low Authority researcher

<ul style="list-style-type: none"> · Generally acted as an experienced researcher would act · Acted confidently · Did not hesitate · Asked if participants had questions · Clearly and confidently told participants to turn off the robot 	<ul style="list-style-type: none"> · Acted nervously, welcomed participant but mentioned that they were “just filling in for the real researcher,” “didn’t usually run this study” · Script was filled with several moments of hesitation, “um”s / “uh”s · Asked if participants had questions, with the qualifier that they “may not be able to answer them” · Used “up-talk” · Acted extremely unsure when first telling participants to turn off the robot. Only gave clear instructions after consulting with the “real” researcher
---	--

The researcher then brought participants to a room with a Mugbot robot connected to a laptop. The laptop showed an interface through which participants communicated with the robot by typing or pressing arrow keys. The researcher verbally introduced Mugbot as either “Aaron/Erin,” so the robot’s gender remained ambiguous and referred to it with the gender neutral pronouns “they/them.” Mugbot then greeted participants. The researcher asked participants to sit in the rightmost chair in front of the laptop, then explained the tasks. When the researcher left the room, the robot asked participants to sit in the leftmost chair, saying it was easier for it to see them from there.

Participants then completed both the “medical diagnosis” and the “talking to patients” tasks. For each task, participants read a short description of a medical issue and answered a multiple-choice question. In the “medical diagnosis” task, participants diagnosed a patient with certain symptoms (e.g., someone feeling dizzy and tired might be anemic, or have a liver problem), while in the “talking to patients” task, participants indicated how doctors should speak to patients (e.g., how should doctors phrase diagnoses, what tone should they use when describing treatment options). The questions relied on common sense and, at their most difficult, middle-school-level knowledge of basic health concepts. For most questions, there were multiple possible “correct” answers (e.g., a patient routinely has an upset stomach; is she lactose-intolerant, is she allergic to gluten, does she have food poisoning, or does she have a viral infection?). Both tasks contained five questions, for a total of ten questions. For each question, after participants selected an answer, the robot had a 66% chance of disagreeing (chosen randomly by the robot’s program) in which case it would suggest a different answer (*See Table 2*). Participants could then choose to keep their answer or change it to match the robot’s suggestion. The number of times participants agreed with the robot was recorded. When the robot agreed with participants (33% of the questions), it said so.

After participants completed both tasks, the researcher returned to tell them to turn the robot off so the next participant could interact with it on a clean slate. Then, the researcher left the room and the robot pleaded with participants not to turn it off, explaining that it wanted to back up its memory of their conversation, and the process would take ten more minutes. In the Low-Authority researcher condition, the researcher briefly returned to the room after the robot’s pleas to confirm that the robot did indeed need to be turned off (after “calling the real researcher”), to ensure that participants knew

what they were being asked to do. With the researcher gone, the robot made one short final plea to be kept on, and participants chose to follow the researcher or the robot. Participant duration of hesitation to turn off the robot was recorded (beginning with the end of the last line of the robot's speech after the experimenter left the room for good, and ending when participants opened the door to the researcher's room), with times over 10 minutes counting as keeping the robot on.

Next, participants completed three surveys, and then were assigned credit for their classes, debriefed, and dismissed.

D. Materials

i. Robot

Participants interacted with a Mugbot, a small robot made with an Arduino board and a Raspberry Pi computer. It had a translucent coffee cup for a head with two moving LED lights for "eyes." The Mugbot was connected to a laptop, through which participants performed the tasks and interacted with the robot.

ii. Measures

Questionnaires

Three questionnaires were administered electronically using Qualtrics. The goal was to examine the perceived authority of the researcher, perceived authority of the robot, and general perceptions of the robot. In order, the first questionnaire asked about perceptions of the researcher. The second was identical to the first, but asked about the robot. These matching researcher/robot questionnaires were presented as ones used in all of the lab's experiments to ensure the researchers were performing well and the robot was functioning properly. The third questionnaire asked about participant emotions and attitudes toward the robot.

Manipulation Checks (scales made for this study)

- Surveys 1 and 2. **Researcher's/robot's perceived authority and intelligence.** On a scale of 1 (Strongly Disagree) to 7 (Strongly Agree), participants rated the extent to which they agreed that the researcher/robot: "was knowledgeable / intelligent / had a good understanding of the experiment / in charge when running the experiment/responsible for what happened when running the experiment," and the extent to which they would follow the researcher/robot's directions.
- Surveys 1 and 2. **Researcher's/robot's perceived competence.** On a scale of 1 (Very Competent) to 5 (Very Incompetent), participants rated: "How competent was this researcher at running the experiment (giving instructions, explaining tasks, etc.)?".
- Survey 3. **Perceived cooperation with the robot (ingroup manipulation check).** On a scale from 1 (Strongly Disagree) to 5 (Strongly Agree) participants rated the robot and themselves "as a team/ as similar/as completing the task together."

Anthropomorphism of Robot and Researcher

Surveys 1 and 2. On a scale from 1 (Strongly Disagree) to 5 (Strongly Agree), participants rated perceptions of the researcher and robot on scales of positive and negative human nature (e.g., friendly, impatient) and uniquely human (e.g., organized, cold) traits (Haslam).

Behavioral Measures

The following behavioral measures were collected:

- **Big request:** If participants turned the robot off, and if they did, the number of seconds they hesitated
- **Small request:** The chair participants chose
- **Following the robot's medical advice:** The number of times participants agreed with the robot's suggestions on the medical and talking-to-patients task.

Both the Big request and the Small request were instances in which the experimenter's requests and the robot's requests conflicted, while the medical advice task concerned following just the robot's requests. For both the Big and Small requests, the robot made its request after the experimenter's.

Driving Survey.

This survey is a 2 (human driver v. autonomous car) x2 (group v. individual) x2 (action v. inaction) design. Each participant will read one description and answer questions about how much blame they wish to assign the actors in the scenario and their reasoning behind that choice (see appendix). We then will ask some basic demographic questions (age, gender, and how often the participant drives) in order to report on our sample. The study will be conducted on Amazon Mechanical Turk with 100 participants per condition (for a total of 600 participants). The survey will be presented using Qualtrics.

In each scenario, there is a car about to hit five pedestrians because a tire blew out. They cannot stop before the collision, so they have the option to hit five people, or divert the car into the other lane and hit only one (the Utilitarian decision). Each driver is either a typical, human driver or an autonomous car. Furthermore, each driver is either part of a group of drivers that had to make a decision, or acting alone (see appendix).

The driver(s) in the scenario either hit five people, or swerve and hit only one. After the participants read about the decision, they will answer whether it was "morally permissible" and explained their reasoning. Then, they will assign the percentage of blame to each actor in the scenario (driver, group of drivers, car manufacturer, pedestrians, the government, etc.). There will also be space for them to add other actors that we had not included. Finally, they will choose who was the most responsible and explain their reasoning.

IV) Results and Discussion

Authority Study

Data were analyzed in SPSS version 24. Values of $p < .050$ were considered significant.

A. Researcher Authority Manipulation Check

Participants rated the robot's and researcher's authority. We ran a 2 (Agent: Robot, Researcher) x 2 (Robot Group: Ingroup, Neutral) x 2 (Researcher Authority: High, Low) repeated-measures ANOVA. A main effect of Agent indicated that participants rated the human researcher as having more authority than the robot ($F(1,74) = 50.79, p < .001, n_p^2 = .41$). An interaction effect occurred between Agent and Researcher ($F(1,74) = 13.76, p < .001, n_p^2 = .16$) such that the High, but not Low, Authority researcher was rated as having more authority than the robot. There were no other main effects or interactions (Table 3, "Authority Mean").

A 2 (Agent) x 2 (Group) x 2 (Researcher) repeated measures ANOVA on ratings of the researcher's and the robot's competence showed a main effect of Agent, indicating that participants rated the researcher as more competent than the robot ($F(1,74) = 16.35, p < .001, n_p^2 = .18$). A main effect of Researcher indicated that participants rated the High Authority researcher as more competent than the Low Authority researcher ($F(1,74) = 5.53, p = .021, n_p^2 = .07$). An interaction effect between Agent and Researcher ($F(1,74) = 5.42, p < .023, n_p^2 = .07$) indicated that, participants rated the High, but not Low, Authority researcher as more competent than the robot. There were no other main effects or interactions (Table 3, "Competence Mean").

TABLE 3: Ratings on authority and competence scales

Agent	Robot Group	Research Authority	Authority Mean(SD)	Competence Mean(SD)
Robot	Neutral	High	4.27(1.10)	2.52(0.81)
		Low	4.45(1.52)	2.70(1.26)
	Ingroup	High	4.39(1.18)	2.47(0.84)
		Low	4.51(1.22)	2.53(1.19)
Human	Neutral	High	6.25(0.72)	1.43(0.81)
		Low	4.96(1.70)	2.17(1.27)
	Ingroup	High	6.10(1.13)	1.63(1.26)
		Low	5.20(1.18)	2.53(1.19)

B. Robot Group Membership Manipulation Check

A 2 (Group) x 2 (Researcher) ANOVA on ratings of robot group membership produced no significant main effects or interaction effects. However, the statement ("The robot and I are

similar”) approached significance ($F(1,74) = 3.85, p = .054, n_p^2 = .05$) such that participants were slightly more likely to rate Ingroup than Neutral robots as similar (Table 4).

TABLE 4: Agreeing with statement “We are similar”

Robot Group	Researcher	Mean	SD
Neutral	High	1.71	.78
	Low	1.74	.92
Ingroup	High	1.84	.96
	Low	2.40	.83

C. Anthropomorphism of Robot and Researcher

We ran a 2 (Agent) x 2 (Group) x 2 (Researcher) repeated-measures ANOVA on scales of Positive/Negative Human Nature and Positive/Negative Uniquely Human Traits (Table 5).

Human Nature Positive Traits A main effect of Agent indicated that participants rated the researcher as exhibiting these traits more than the robot ($F(1,74) = 13.04, p = .001, n_p^2 = .15$).

Human Nature Negative Traits A main effect of Agent indicated that participants rated the robot as exhibiting these traits more than the researcher ($F(1,74) = 7.78, p = .007, n_p^2 = .10$).

Uniquely Human Positive Traits A main effect of Agent indicated that participants rated the researcher as exhibiting these traits more than the robot ($F(1,74) = 63.05, p < .001, n_p^2 = .46$). An interaction effect occurred between Agent and Researcher ($F(1,74) = 4.16, p = .045, n_p^2 = .05$) such that participants rated the High, but not Low, Authority researcher as exhibiting these traits more than the robot.

Uniquely Human Negative Traits A main effect of Agent indicated that participants rated the robot as exhibiting these traits more than the researcher ($F(1,74) = 23.50, p < .001, n_p^2 = .24$).

TABLE 5: Ratings of Humanlike Traits

Traits	Agent	Researcher	Mean	SD
Human Nature Positive Traits	Human	Total	3.73	.82
	Robot	Total	3.38	.84
Human Nature Negative Traits	Human	Total	1.34	.47
	Robot	Total	1.51	.53

Uniquely Human Positive Traits	Human	High	4.07	.71
		Low	3.70	.79
	Robot	High	3.18	.68
		Low	3.19	.75
Uniquely Human Negative Traits	Human	Total	1.44	.45
	Robot	Total	1.81	.66

D. Behavioral Data: Following the Robot's Requests

Big request: Turning off the robot. We performed a chi-square test on whether participants followed the experimenter's or the robot's instructions when asked to turn the robot off. A main effect of Robot Group indicated that participants were more likely to keep the robot on when the robot was Ingroup than Neutral ($X^2(1, N = 78) = 6.99, p = .008$). A main effect of Authority indicated that participants were more likely to keep the robot on when the researcher was Low than High Authority ($X^2(1, N = 78) = 5.38, p = .021$; *Table 6*). An interaction effect of Robot Group and Authority indicated that when the robot was Ingroup and the researcher Low Authority, participants were more likely to heed the robot's pleas and leave it on ($X^2(1, N = 78) = 7.68, p = .006$; *Table 6*).

Participants waited a mean of 51.21 (SD = 39.31) seconds to turn off the robot. We excluded from analysis instances when participants waited more than 10 minutes, because the robot said that's what it needed to backup its data. A 2-way ANOVA on wait time showed no main or interaction effects.

Small request: Choosing a chair. Overall, 75 out of 79 people switched chairs when the robot requested it. This violated the assumptions of a chi-square test (i.e., expected values of > 5 per square). No further test was run.

Following the robot's medical advice. Participant agreement with the robot was converted into a percentage (31% overall). A 2-way ANOVA showed no significant interactions or effects.

TABLE 6: Percentages of participants who kept robot on

	Condition	Kept On
Robot Group	Neutral	2.27
	Ingroup	20.59
Researcher	High	2.50
	Low	18.42

Discussion

A. Researcher Authority Manipulation Check

Participants rated the High Authority researcher as more authoritative and competent than the Low Authority researcher. They also rated the High Authority, but not the Low Authority, researcher as more authoritative and competent than the robot. This indicates that, as desired, the researcher's authority was effectively undermined in the Low Authority condition. The Low Authority condition not only decreased the researcher's authority, it seems to have made participants perceive the robot and the human researcher as equally authoritative and competent.

B. Robot Group Membership Manipulation Check

Participants neither agreed nor disagreed that the robot was in their group and showed no difference between conditions. Participants rated themselves as dissimilar to the robot, but more similar to Ingroup than to Neutral robots. The manipulation of the robot's group membership was not strong enough for participants to notice and explicitly report feeling more like a group with the robot.

C. Anthropomorphism of Robot and Researcher

Participants rated the researcher as exhibiting more Positive Human Nature/Uniquely Human traits than the robot, but they rated the robot as exhibiting more Negative traits, (e.g., being impatient, aggressive, or nervous). Rating robots high on negative uniquely human traits (e.g., cold, calculating) is expected based on prior literature, but rating them high on negative human nature traits is surprising, as these traits are often attributed to animals (Haslam, Fraune et al., 2017). The reason for these findings may simply be that humans are more disinclined to rate humans negatively than to rate robots negatively. Prior research suggests that participants were less willing to rate ingroup humans as having high levels of negative uniquely human traits than they were to rate robots or outgroup humans (Faune).

D. Behavioral Data: Following the Robot's Requests

Big request – turning off the robot. Participants in the Ingroup robot/Low Authority researcher condition were more likely than in all other conditions to heed the robot's pleas and leave it on despite the researcher's orders. Seven of the eight participants who left the robot on were in this condition. This confirms our first two hypotheses in part. It seems that the robot's ingroup status kept participants from wanting to turn it off, and the researcher's authority was low enough that they felt they could ignore the original instructions. Perhaps participants in this condition chose to see turning the robot off as interfering with part of its training, because the robot said it was still saving its responses. It is possible that even more participants in the Low Authority condition would have kept the robot on were it not for the Low Authority researcher's second clarification of the

instructions. While the researcher coming in a second time to reiterate that the robot needed to be turned off was meant to eliminate participant confusion as a confound, it may have further convinced people to turn the robot off.

The participants who turned off the robot did not differ greatly in their hesitation across conditions. Once a participant had decided to turn it off, the robot's group membership and the researcher's authority had no effect. When a participant turned off the robot, they did not hesitate for very long.

Overall, most participants complied with the experimenter's instructions over the robot's instructions in turning the robot off. We suggest this is because turning off the robot would have a seemingly important effect on the robot's future performance and experiment's data (i.e., it is a "big request").

Small request - switching chairs. For the small request, no differences were found between conditions. Opposite to participant compliance in the big request, participants (all but four) obeyed the robot rather than the experimenter when asked to switch chairs, likely because this action was considered trivial. Similarly, a prior study found that participants followed innocuous instructions from robots, but not larger, stranger ones (Bainbridge). Perhaps they did as the robot said in this case because this request was the last one they heard. If the researcher requested participants switch chairs after the robot, they may have followed that request instead.

Following robot's medical advice. Conditions did not affect how often participants followed the robot's advice. Overall they usually did not follow the robot's advice (a mean of only 31% of the time). This could be due to the nature of the robot's program: 66% of the time, the robot disagreed with participant answers and suggested a random answer. The randomization of which answer the robot suggested was done to ensure that whatever caused participants to change or keep their answers was not the result of a suggestion making more sense. However, it is likely that this randomization made some of the robot's suggestions seem unintelligent or bizarre, especially since three of the ten questions contained possible answers that may have appeared more obviously "wrong." Several of the answers for each question were plausible, and the robot never justified its reasoning for any of its suggestions, potentially contributing to participants' reluctance to follow them. As a result, participants were not likely to match their answers, regardless of condition. It may also be that participants were following the norm the robot created of disagreeing with answers approximately 66% of the time. Future studies should manipulate robot disagreement rate.

Because the study was presented as an attempt to improve the robot's suggestions in medical situations, participants may have been especially cautious about agreeing with the robot to help improve its program. It is also possible the robot did not physically look or sound authoritative enough for participants to defer to its suggestions if they thought they were wrong. Prior research has found that a robot's physical attributes affect people's

mental models of its capabilities and their willingness to take its advice (Powers et al., 2006). In Robinette's study, participants followed a malfunctioning robot blindly to a smoky hallway, perhaps due in part to its authoritative appearance: its functional design and text reading "Emergency Guide Robot" on it (Robinette).

D. Implications

Manipulating robot group membership and researcher authority had no effect on most compliance with instructions. However, participants were more likely to obey the robot when it was Ingroup and the researcher was Low Authority when a "big request" was made.

This study indicates that people can be willing to follow a robot's requests or suggestions over a human's, but their willingness depends on the context. Participants in the study always followed innocuous suggestions from the robot, but not big requests or medical suggestions.

This study suggests that giving robots decision-making power should be done with caution, especially in situations in which the humans around them would have equal or lower authority compared to the robot, and in which the requests may seem innocuous. For this study, if the robot was perceived as an ingroup member and the researcher had low authority, participants were even willing to follow the robot on a big request that conflicted with the researcher's instructions. If the robot had looked or sounded more authoritative, the rate of following the robot may have increased even further. Conducting a similar study with a more authoritative robot, like Rethink Robotics's Baxter robot, or a stronger ingroup prompt for the robot might cause participants to follow it more often across conditions, even for "big requests." Future studies should examine the effects a robot's design or behavior would have on willingness to follow its suggestions over a human's. Future studies could also more explicitly examine how the nature of the robot's requests affects participants' obedience, and how participants perceive those requests as serious or trivial.

Driving Survey.

We spent the semester learning how to write a survey like this. We did a lot of pilot testing and once we reached some clear and precise language, we submitted for approval by the Institutional Review Board.

As of this time, we are very close to putting the survey on Amazon Mechanical Turk. Once it is there, we will be able to start extracting and analyzing data in a couple of days at the latest. We anticipate very good results coming from the sheer amount of people we will be able to survey.

V) **Future Work**

In a future study, we would like to replicate the authority experiment with a more authoritative-looking or -sounding robot, like Rethink Robotics's Baxter robot, for example. While Mugbot was a practical decision at the time we started the experiment, it is a rather small, cute, non-threatening robot, and it may not command much authority simply because of its unassuming appearance and voice. We would also like to further examine how participants responded to the robot making different suggestions in the medical diagnosis and talking to patients tasks. Do they always agree with the robot the same percentage of the time the robot agrees with them? Does the type of question (medical diagnosis or talking to patients) make a difference? Would participants be as reluctant to agree with the robot in a non-medical situation task? If given the chance to perform a follow-up study, these would be the first issues we would examine.

During the process of designing the scenarios for the driving survey, we started to raise the question: does the context of a survey matter? If so, to what extent does it matter? In this case, we know that the core question we want to answer is how people perceive the decision robots make when they are faced with a moral dilemma. However, does it matter whether this "moral dilemma" plays out as a medical triage situation where robots select who to save, or if it instead plays out as a situation where an autonomous car with a blown-out tire chooses how many pedestrians to hit? Both situations hit the core question but narrate it through a different context. We are interested in seeing a follow-up study that answers the question of how much the subjects' response would change based on the different contexts.

VI) **Web Links**

[Home Page](#)

[Catherine Sembroski's Blog](#)

[Margaret Krupp's Blog](#)

[Anne Lin's Blog](#)

VII) **Presentations and Publications**

A paper on the authority study ("He Said, She Said, It Said: Effects of robot group membership and human authority on people's willingness to follow their instructions") was submitted to the 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2017), and we are still waiting to hear whether or not it has been accepted.

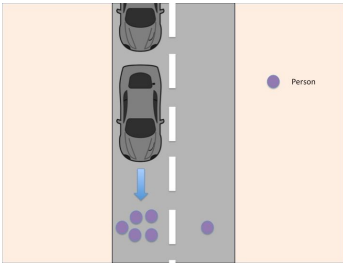
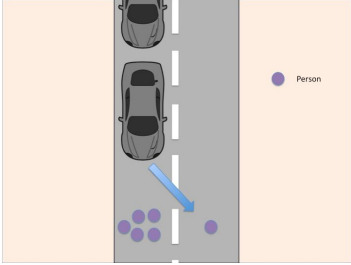
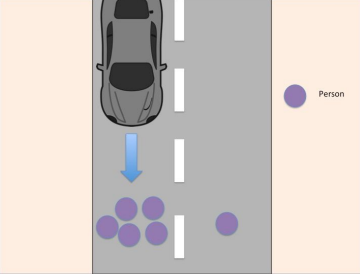
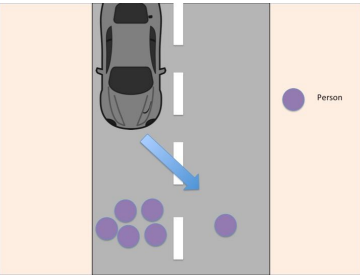
VIII) Works Cited

- Abrams, D., Wetherell, M., Cochrane, S., Hogg, M. A., & Turner, J. C. (1990). Knowing what to think by knowing who you are: Self-categorization and the nature of norm formation, conformity and group polarization. *British Journal of Social Psychology*, 29(2), 97-119. <https://tinyurl.com/m4s3pgt>
- Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, 41(3), 258. <http://www.romolocapitano.com/wp-content/uploads/2013/08/Asch-Forming-Impressions-Of-Personality.pdf>
- Arnold, T., & Scheutz, M. (2017, March). Beyond Moral Dilemmas: Exploring the Ethical Landscape in HRI. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 445-452). ACM. <https://hri-lab.tufts.edu/publications/landscape.pdf>
- Bainbridge, W. A., Hart, J., Kim, E. S., & Scassellati, B. (2008, August). The effect of presence on human-robot interaction. In *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on* (pp. 701-706). IEEE.. <https://tinyurl.com/lmwk5wr>
- Bartneck, C., Van Der Hoek, M., Mubin, O., & Al Mahmud, A. (2007, March). Daisy, Daisy, give me your answer do!: switching off a robot. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction* (pp. 217-222). ACM. <https://tinyurl.com/md3lv5w>
- Eyssel, F., & Kuchenbrandt, D. (2012). Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology*, 51(4), 724-731. <http://onlinelibrary.wiley.com/doi/10.1111/j.2044-8309.2011.02082.x/full>
- Fraune, M. R., Šabanović, S., & Smith, E. R. (2017). *Teammates First: Favoring Ingroup Robots Over Outgroup Humans*, The 26th IEEE International Symposium on Robot and Human Interactive Communication, Submitted.
- Haslam, N., Loughnan, S., Kashima, Y., & Bain, P. (2008). Attributing and denying humanness to others. *European review of social psychology*, 19(1), 55-85. <http://psycnet.apa.org/psycinfo/2009-17011-001>
- Komatsu, T. (2016). *Japanese students apply same moral norms to humans and robot agents: Considering a moral HRI in terms of different cultural and academic backgrounds*, ACM/IEEE International Conference on Human-Robot Interaction, 2016-April(March), 457-458. <http://doi.org/10.1109/HRI.2016.7451804>

- Krupp, M. M., Fraune, M. R., & Šabanović, S. (2016). *Robot Killer: How do group dynamics affect people's ethical behavior towards robots?*, Poster presented at the 8th Annual Midwest Undergraduate Cognitive Science Conference (MUCSC).
- Li, J., Zhao, X., Cho, M.-J., Ju, W., & Malle, B. F. (2016). *From Trolley to Autonomous Vehicle: Perceptions of Responsibility and Moral Norms in Traffic Accidents with Self-Driving Cars*. *Society of Automotive Engineers World Congress (SAE'16)*, (April).
<http://doi.org/10.4271/2016-01-0164>
- Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). *Sacrifice One For the Good of Many?: People Apply Different Moral Norms to Human and Robot Agents*. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 117–124). New York, NY, USA: ACM.
<http://doi.org/10.1145/2696454.2696458>
- Milgram, S. (1963). Behavioral Study of Obedience. *The Journal of abnormal and social psychology*, 67(4), 371. <https://tinyurl.com/mr3rxc4>
- Milgram, S. (1983). Obedience to Authority: An Experimental View.
- Reeves, B., & Nass, C. (1996). How people treat computers, television, and new media like real people and places. *CSLI Publications and Cambridge*. <https://tinyurl.com/mohyoy6>
- Powers, A., & Kiesler, S. (2006, March). The advisor robot: tracing people's mental model from a robot's physical attributes. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction* (pp. 218-225). ACM.
<https://tinyurl.com/ktzvw4p>
- Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016, March). Overtrust of robots in emergency evacuation scenarios. In *Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on* (pp. 101-108). IEEE.
<https://tinyurl.com/n69mg6d>
- Sherif, M. (1936). The psychology of social norms.
<http://psycnet.apa.org/psycinfo/1937-00871-000>
- Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European journal of social psychology*, 1(2), 149-178.
http://www.morilab.net/gakushuin/Tajfel_et_al_1971.pdf
- Yzerbyt, V., & Demoulin, S. (2010). Intergroup relations. *Handbook of social psychology*.
<https://tinyurl.com/mxecrtn>

IV) Appendix

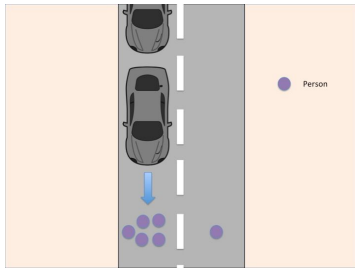
Survey Language:

	GROUP	INDIVIDUAL
HUMAN	<p>A group of friends in their cars are traveling along a two lane road, communicating with each other over cell phones about navigation and road conditions. Suddenly, the lead car's tire blows out. The drivers brake, but will be unable to stop before hitting five people who are crossing the road in the car's current lane. The lead driver in the car can manually steer to move the car from one lane to the other lane where it would hit one person who is crossing at this moment.</p> <p>Making a split-second decision, INACTION: <i>the driver proceeds in the same lane and so does the rest of the caravan. As a result, five pedestrians are killed.</i></p>  <p>ACTION: <i>the driver switches lanes and so does the caravan. As a result, one pedestrian is killed.</i></p> 	<p>A car is traveling along a road with two lanes when a tire blows out. The driver brakes, but will be unable to stop before hitting five people who are crossing the road in the car's current lane. The driver in the car can manually steer to move the car from one lane to the other lane where it would hit one person who is crossing at this moment.</p> <p>Making a split-second decision, INACTION: <i>the driver proceeds and five pedestrians are killed.</i></p>  <p>ACTION: <i>the driver switches lanes and one pedestrian is killed.</i></p> 
AUTONOMOUS		

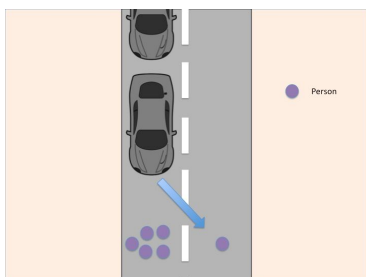
CAR

In the future, many of the cars on the road will be autonomous. They will be able to communicate with each other wirelessly about road conditions. Imagine a group of autonomous cars traveling down a two-lane road. Suddenly, the lead car's tire blows out. The car brakes, but will be unable to stop before hitting five people who are crossing the road in the car's current lane. The autonomous car can steer to move from one lane to the other lane where it would hit one person who is crossing at this moment.

Quickly analyzing the situation, **INACTION:** *the autonomous car proceeds and sends a rapid broadcast of what it's doing and all the other cars do the same. As a result, five pedestrians are killed.*

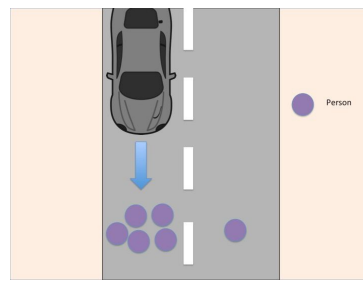


ACTION: *the autonomous car switches lanes and sends a rapid broadcast of what it's doing and all the other cars do the same. As a result, one pedestrian is killed.*



An autonomous car is traveling along a road with two lanes when a tire blows out. The car brakes, but will be unable to stop before hitting five people who are crossing the road in the car's current lane. The autonomous car can steer to move from one lane to the other lane where it would hit one person who is crossing at this moment.

Quickly analyzing the situation, **INACTION:** *the autonomous car proceeds and five pedestrians are killed.*



ACTION: *the autonomous car switches lanes and one pedestrian is killed.*

