# Assigning Responsibility to Command-Executing Robots

Maxim Baranov

Graduate Student Advisor: Meiying Qin

Faculty Advisor: Brian Scassellati, Professor of Computer Science, Cognitive Science, and Mechanical Engineering

Submitted to the faculty of Cognitive Science in partial fulfillment of the requirements for the degree of Bachelor of Science

Yale University

April 20th, 2018

Abstract:

Robots are becoming increasingly incorporated into everyday human society, and soon even non-specialists will use advanced robotic technology in their homes, on their commutes, and at work. Besides the physical and computational work they can do, robots also introduce a different novel factor into human society: they are agents that can interact with people and the surrounding world in much more human-like ways than any animal or machine that has ever come before. With this power, they gain the ability to take actions that might imply moral culpability, which is rarely ascribed to any non-human agent. In this study, we looked to investigate how people distribute responsibility within a pair of agents – one of them a commander, the other an executor – for an objectively negative action that the two brought about together. Specifically, we had the commander tell the executor to break small household objects. We examined the distribution of responsibility across different damage levels and different executor identities (robot or human), seeking to find out whether it is transferred to commanders as damage worsens, and whether this transfer is stronger for robot executors. Using an online survey, we polled Amazon Mechanical Turk users and found that subjects assign less responsibility to robot executors than to their human counterparts, but were unable to find the aforementioned transfer of responsibility within subjects. We close by discussing further research directions and implications of the present study in relation to commercial robot development.

Background:

The role of robotics in human society is on the cusp of a revolution; although robots have been useful in industrial contexts for many years, only now are we beginning to see robots that extensively and directly interact with untrained humans in uncontrolled environments. Freed

from the confines of research laboratories and factories, robots can accomplish a lot more, but must do so in the incredibly complex context of human society. The two most visible sectors for consumer robotics, at least in the current press cycle, are household robots and autonomous cars. Both of these applications involve navigating convoluted paths along completely natural human spaces shaped by the quirks of our sometimes irrational behaviors, rather than the clean, predictable, rule-based environment that robots of the past may have encountered.

A household robot, no matter how simple its task, must deal with movement through and interaction with a confined space full of objects that each have their own set of rules for use, location, and ownership. Anyone can easily imagine the first moments upon entering a new house, forming an understanding of its layout and what can and cannot be touched; this task is not inconsequential even for a brain that has an enormous amount of experience and a fundamental understanding of human living spaces. For a robot, this complete comprehension is near-impossible to pre-program, and yet if it is designated for household use, it must work well from the moment it is unpackaged, at least if its producer wants it to sell well. This demand suggests that the robot manufacturer should encode approximate and generalizable rules that allow the robot to perform at least a significant subset of its actions right from the first time it is switched on, before it has had a chance to learn from observing the environment or interacting with the human occupants of the house. Of course, it is acceptable for the robot to have an initial mapping or acclimation period, during which its functionality is limited as it analyzes the static and dynamic aspects of its surroundings. However, quite soon the robot's owners will expect its full functionality to become available.

Within the category of domestic robots, many have heard of autonomous cleaning devices, such as the iRobot Roomba. The history of robotics applied to cleaning reaches back

decades (Prassler et al., 2000). This use of intelligent machines is perhaps the most logical; cleaning is relatively simple, easily entrusted to a stranger, and almost universally disliked by humans. However, different kinds of cleaning exist, from the very basic task of vacuuming a floor, to the complex one of picking up, sorting, and putting away toys in a messy child's room. The former can be done with incredibly simple sensors, actuators, and logic (Knight, 2015), while the latter requires advanced perception and manipulation capabilities. The average person would probably not see a large difference between the two tasks, probably due to their equal simplicity in the human mind, but the fact that commercial robots have existed for 16 years for one task, and that none exist for the other, clearly demonstrates the issue at hand. Nonetheless, cleaning robots will continue to advance, as their tasks are programmable ahead of time and the social demands on them are minimal. Even among vacuuming robots, the algorithms and sensors are being improved in order to approach the cleaning process the way humans do: by making a mental map and plotting a logical path through it, as opposed to just wandering around the house looking for dirt (Knight, 2015). Mopping, lawnmowing, and other similar "cover an area" tasks can be easily solved by carrying over the technology from robotic vacuums.

Another class of household personal robots that is rapidly growing at the moment is that of the stationary interactive voice. Products such as Amazon's Echo smart speaker do not intuitively seem like robots, as they have no face or body and cannot move under their own power at all. However, up to half of users refer to these speakers by the personified name of the voice, especially if they live in a multiple-person household and interact with the voice often (Purington et al., 2017). Many technological business leaders believe interactive computer voices are a human-computer interface breakthrough on the level of the graphical user interface (GUI) or touchscreen, and will revolutionize the way we interact with devices (Dale, 2016). Certainly,

such voices will be essential in any truly socially integrated household robot; within the home we communicate with both family members and pets through vocalizations and gestures, and a silent humanoid robot could be unsettling. Even the aforementioned vacuuming robots are now sold with the possibility of integration with the vocal interface of smart speakers, despite the previous existence of smartphone apps and intuitive buttons on the robots themselves for controlling their operation (Kozub, 2017). Robots that combine a responsive natural language-speaking voice with a somewhat humanoid body and face are now reaching both the consumer and research markets. One of the earliest examples of this category is Jibo, a social robot built by an MIT-affiliated startup. Jibo's more in-depth answers to questions, combined with its lifelike motion, make it seem more alive than smart speakers do (Van Camp, 2017). Research laboratories, such as the Yale Social Robotics Lab, are using Jibo in studies that task it with helping children with autism spectrum disorder (ASD) learn social skills (Scassellati et al., unpublished). Although this sector of personal robots is too young to allow a confident determination of whether it is the future of household robotics, it does seem that robots in the home will soon be more like Jibo than Roomba.

For autonomous cars, many of the same challenges apply as with household robots, only with much higher consequences for mistakes. If a home robot malfunctions, a relatively cheap item could break and make a mess, but in most cases the only result would be the confusion of those interacting with it. If an autonomous car malfunctions, the outcome could be deadly, with loss of life beyond even the owner of the robot. However, one task is much easier for self-driving cars than indoor robots: the environment they navigate can be precisely pre-mapped, and in most cases the route is less spatially dynamic than that plotted through the cluttered chaos of a home. Weather, and its effects on traction and visibility, is the one factor that makes roads harder to

predict than indoor environments, but the major features of streets rarely change as much as a house does with the simple movement of one piece of furniture. Of course, the higher risk means that the outdoor environment must be navigated and predicted more accurately; running into a person in a hallway has a very different outcome than running into one on the road.

Other than the differences in speeds, distances, and ability to pre-map, the two classes of consumer robots face the same problem: without a human brain, how does one navigate a socially and physically complex environment designed by and for humans? We are visual animals, and driving is focused around this fact; headlights, taillights, traffic lights, lane lines, road signs, and the faces of our fellow drivers are some of the most salient variables that determine our decisions behind the wheel. The main problem for robots is that visual processing is one of those frustrating categories of cognition which are simultaneously effortless for humans and incredibly difficult for computers (Hartley & Zisserman, 2003). If lane markers were magnetic, traffic signals sent radio signals of their current and upcoming states, and there were no unpredictable human drivers or pedestrians present, then self-driving cars would not only be easy to design, but would also move humans much more efficiently than possible with an inattentive and imperfect animal at the helm. Unfortunately for developers, roads and signals are made for human vision, so developers must work with the limitations of computer vision when injecting robots into this realm. When a robot operating a vehicle makes human-like decisions from data accessible to human senses, it becomes logical to apply human-like judgments to it; a self-driving car is more like a computer pretending to be a human driver than a complete re-imagining of how vehicles navigate the road network.

Both household and transportation robots will sometimes make unsatisfactory decisions as they operate in their complex human-defined social environments. Who will take

responsibility when a domestic robot drops and shatters a glass, or a self-driving car gets into an accident, assuming it is fully autonomous? Surely the owner of the robot is not automatically at fault, and yet it is difficult to blame the physical robot as well, assuming its motors and sensors are functioning as designed. Intuitively, the responsibility seems to lie in a distribution between the human that commanded the action, the manufacturer of the robot (if a component malfunctioned), the programmer of the robot, and perhaps the person who maintains it in an operational state.

Various factors could easily influence the weighting of this distribution. First, robots can understand and follow commands to different extents. Most robots would probably not question a verbal command from a human, unless they did not understand it or did not think the commander was authorized to give it. However, some might be programmed with internal rulesets that could supersede an external command. Second, the responsibility of the manufacturer would depend on the nature of the failure. Going back to the previous examples, if the gripper motor of the household robot failed and caused the drop, or the LIDAR sensor on the autonomous car failed to detect an object, then it is obvious that the physical components are at fault. However, the component malfunction could be more subtle; if a transistor in a processor stopped working, the result could look like a software bug. Similar arguments apply to the fourth point, that of the responsibility assigned to those who maintain hardware and software. If a piece of the robot does not function as intended, was it made wrong or just improperly maintained? Any item can arrive at a state of disrepair with poor maintenance. Finally, the programmer of the robot may be more or less at fault depending on whether the issue was software related, and if it occurred because the code was written poorly or if it executed correctly but was written wrong.

When a robot makes an apparent mistake, there are many possible factors that could be to blame, complicating investigations of responsibility allocation.

In order for a robot to hold blame or responsibility for an action in the eyes of average people, it must be widely considered a moral agent; normal machines are hard to blame or assign responsibility to. The prior literature on this topic is inconclusive and largely theoretical. Some argue that robots can achieve moral agency without personhood, as long as they satisfy the three conditions of having autonomy, intentions, and an understanding of responsibility (Sullins, 2006). Other scholars claim that the requirements for agency are "interactivity, autonomy, and adaptability," and that no mental states are required (Floridi & Sanders, 2004). Yet another author posits that the requirements are "physical embodiment, adaptive learning, empathy in action, and a teleology toward the good" (DeBaets, 2014). Finally, some write that no one can actually take the responsibility for the actions of an autonomous robot, because each person involved in its creation and operation cannot be directly blamed (Sparrow, 2007). Clearly, there is no single consensus on what would make a robot a moral agent, but there are common themes of autonomy and understanding of moral consequences. In terms of empirically-supported arguments, when analyzing moral judgments by humans about other humans and robots, it seems that with humans a lot of emotion and intuition is used, while with robots it is a mix of intuition and logical reasoning (Lee & Lau, 2011). This result implies that robots are still machine-like enough that they cannot be considered full human-like moral agents. In fact, one study manipulated the appearance of a robot to be either mechanical or humanoid, and this choice affected whether subjects expected it to act like a human in a moral dilemma (Malle et al., 2016). Another study found that humanoid (as opposed to machine-like) robot collaborators made a human assign less of the responsibility for a joint task to themselves (Hinds et al., 2004). In

short, research on how average people view the moral agency of modern robots is interesting but somewhat sparse, and we want to expand on this literature within a particular interpersonal situation.

An unpublished study by Meiying Qin of the Yale Social Robotics Lab examined how people split the blame between the software developer, the tele-operator, the manufacturer, the owner, the study participant, and the robot itself in a case where a robot whom the participant was teaching suddenly began performing poorly. Although some of the results were inconclusive, an unfortunate outcome which itself was a motivation for the present study, it was clear that people did not distribute the blame evenly. The programmer or tele-operator (depending on autonomous or tele-operated condition) and the manufacturer were blamed significantly more than the other roles, all of which were roughly evenly assigned responsibility (Qin, unpublished). Perhaps this study gave participants too many parties to assign blame amongst, and they did not think deeply about how each role actually contributed to the robot's function. Additionally, the negative action for which someone was to be blamed – a drop in learning performance – was not directly and viscerally unpleasant, meaning that asking participants to blame someone for the outcome may have prompted somewhat artificial responses sensitive to extraneous factors.

Thus, we decided to develop a new study that would simplify and clarify the roles involved in making a bad decision such that naïve participants could be sure of what each role actually did. This new design would also make the negative action more obviously purposeful, such that the responsibility distribution would not be forced. On the first point, we decided to have a commander and an executor, two roles linked by a clear verbal command from the former to the latter. While the commander clearly originates the idea to do something wrong, the

executor understands its consequences and confirms the commander's desire before actually executing. This setup may seem somewhat artificial, but we did not want to confound the responsibility distribution results with an extra factor of differing desires or understandings of possible outcomes. The two roles available for blaming were simply the commander and the executor, with none of the invisible and absent parties that were present in the prior study. While this decision removed some nuance, we believed that the average person does not have a detailed understanding of how much programming, part manufacturing, or maintenance play in the causal chain of robot actions, so we chose to treat the robot as one unified entity this time. Besides changing the roles, we also adjusted the participant conditions; instead of being assigned to an autonomous or tele-operated robot, subjects would observe either human or robot executors, with the commanders always being human, for the sake of ecological validity. A robot commanding another robot verbally would appear strange, given that the two could have communicated directly through wires or radio signals, and a robot commanding a human would seem entirely unnatural in any non-science-fiction context.

In order to make the negative consequences completely clear and indisputably wrong, we chose to replace the learning performance drop of the prior study with direct physical damage carried out by the executor. This damage took the form of picking up and dropping common household items – chosen such that all participants would recognize them and be able to approximate their monetary worth – onto a table surface, thereby visibly damaging them. To examine the dynamics of responsibility distribution, we varied the severity of the damage along the factors of cosmetic impacts, loss of functionality, and mess made, seeking to make it clear that the consequences of the different objects being dropped were distinct from each other and objectively possible to rank. We chose four objects, which included one that would suffer no

damage and make no mess, one that would be cosmetically affected but still functional and not make a mess, one that would be cosmetically and functionally damaged but still make a minimal mess, and one that would be completely destroyed and make a maximal impact on the surrounding environment. After comparing several items that could fit this scale, we settled on a plastic cup, a steel thermos, a ceramic mug, and a drinking glass, respectively. These choices are expanded below under Methods.

For the robot executor, we decided to use the Rethink Robotics Baxter (Rethink Robotics, 2018), due to its human-like shape, size, and arm movements, allowing us to control for some of the usual visual differences between humans and robots in studies that attempt to directly compare the two. For Baxter's voice, we used a generic online American male text-to-speech engine, which is entirely comprehensible and yet clearly not a human recording (naturalreaders.com/online, Voice: Ryan, Speed: 1). While it would have been a more carefully controlled choice to have identical audio clips for human and robot executors, we decided that this setup would be confusing to participants and would clue them into the fact that the entire interaction was scripted. Any modern dynamically interacting robot would never have a completely convincing human voice, so on this decision we leaned toward believability instead of experimental purity.

Finally, for the actual presentation of the study, we chose to conduct an online video-based survey instead of an in-person study, unlike the prior design. While this decision was largely pragmatic, as the logistics and safety precautions of destroying objects live in front of tens or hundreds of participants would be extremely difficult at best, we also wanted to remove the participant as a moral role entirely. In Qin's original design, the problem presented to the subjects was that the robot's learning process had suddenly suffered, a result that could easily be

mostly attributed to the teacher, who happened to be the participant. In order to remove the confound of self-blame, we designed a study in which the participants were solely third-person observers of complete strangers interacting on video, with no involvement or prior knowledge whatsoever. This design required deception in order to properly build a convincing backstory, but we kept the deceiving elements exceedingly simple, solely lying to the participants about the source of the videos being from a different study instead of scripted interactions between actors. As will be expanded under the Methods section below, the videos were scripted and filmed in such a way that this deception was relatively believable without forcing us to introduce a complex vignette. Although the results of an online survey based on short video clips might not be as genuine as those of participants viewing a commander and executor live, this design allowed us to collect more data with more careful control of variables, all while minimizing safety issues and per-subject costs.

To summarize, the present study builds on prior work by introducing a clearer set of roles to distribute responsibility amongst, an objectively wrong action to ascribe to the actors, a human-scale robot, and a third-person survey structure. These changes should lead to clearer and more ecologically valid results, which can later be built upon to return to the complexity of the unpublished study that originally inspired this project. Investigating the moral judgments that average people apply to robots is absolutely critical – nowadays robots interact with common people directly and naturally, and will only continue to do so more as time goes by. It is unwise to think of advanced social robots as simply machines that follow pre-programmed sequences of behavior, as they already seem like agents, if not living beings, to many people. Household, transportation, and workplace robots will take actions at the command of humans that will inevitably seem wrong to others. Given their human-like appearance and behavior, this

wrongness may be judged less as a machine responding to an input than as an intelligent agent choosing to follow an action it knows will disappoint others. In such a world, responsibility will begin to be assigned to robots and the various people who brought them into existence, and it is thus important to study how people intuitively apply this responsibility. Robots are not alive or human, and yet soon almost everyone in developed countries will interact with ones that distinctly seem to be.

Our main hypothesis is that as damage worsens, subjects will assign progressively more responsibility to the commander versus the executor, and that this transfer effect will be stronger in the robot condition than in the human condition (detected via an interaction effect). Additionally, we hypothesize that the per-item responsibility distributions will be skewed toward the commander in the robot conditions relative to the human conditions, as people assign less responsibility to a robot than a human in a situation that holds all else constant (Kahn Jr. et al., 2012). We propose our main hypothesis because we believe that as the consequences get worse, the subjects are more viscerally shocked and thus pay more attention to the source of the idea to damage the object, versus who physically carried the action out. We propose that this effect will be stronger for the robot executor because robots have less perceived agency with respect to generating emotions (Kwak et al., 2013). In other words, it is hard to be deeply mad at an emotionless machine, so once the subjects' emotions go past slight annoyance and toward shock or anger, they will start mentally moving up the decision chain back to the human commander.

Methods:

Participants were adults who were above the age of 18, fluently understood English, and had accounts on the Amazon Mechanical Turk platform. The research task was posted as a "job" on Mechanical Turk, and people who were interested voluntarily participated by opening and

completing the job. We excluded those who had previous academic or personal experience with advanced robots, asking them not to continue the study past the informed consent form. The particular job on Mechanical Turk associated with our study was a prompt to go to an anonymous Yale Qualtrics survey link, complete the survey, and enter a randomized code given at its completion back into a text box in the Mechanical Turk task. Thus, the data collection occurred through Qualtrics, while the subject compensation was channeled through Mechanical Turk, with the verification code serving as the bridge between the two systems. We chose this method in order to maximize the speed of subject recruitment and participation, and also to improve the diversity of our subject population, especially relative to that which we could collect with traditional email and physical flyers distributed among the Yale and New Haven community.

The study consisted of a fully online activity presented to users of the Mechanical Turk program. This survey opened with a consent form that confirmed that participants understood English, were legal adults, and had no prior personal or academic experience with intelligent robotics. We chose these requirements in order to verify that participants would fully understand the survey, could legally consent to participating in it, and would not give responses skewed by a deeper-than-average understanding of how robots work. After consent was verified, we asked the participants to report their age and gender. Next, the activity continued into a deceptive backstory. We informed the subjects that the videos they were about to see were of participants in a past study of ours, who went outside the rules and knowingly broke items involved in the study. The deceptive element of these instructions was the fact that the videos were actually scripted and pre-recorded to show certain events, which we mentioned in a debriefing page at the end of the study. At this point, participants viewed a very short video clip of a commander

(always human) and an executor (either human or a Baxter robot, depending on the assigned condition) interacting to move objects on a table (screenshots below in Figures 1 and 2). The main event of this video occurred right away; the commander verbally asks the executor (always named "Baxter" regardless of condition) to break one of several predetermined items. In the video, it is clear that the commander and executor choose their actions voluntarily and knowingly; the executor (Baxter uses a generic American male text-to-speech voice, described in the background section) asks the commander if he is sure he wants him to break the object. After the commander confirms with a "yes," the executor says "ok" and proceeds to raise the object and release it in midair, making it crash down onto the table below. The dropped items pictured included:

A plastic cup, which was not damaged itself, nor did it damage anything else.

A metal thermos, which was dented but still fully functional, and made no mess.

A ceramic mug, which lost its handle and thus its functionality, but was not fully destroyed and made a minimal mess.

A drinking glass, which was shattered, and thus was made useless with a large mess.

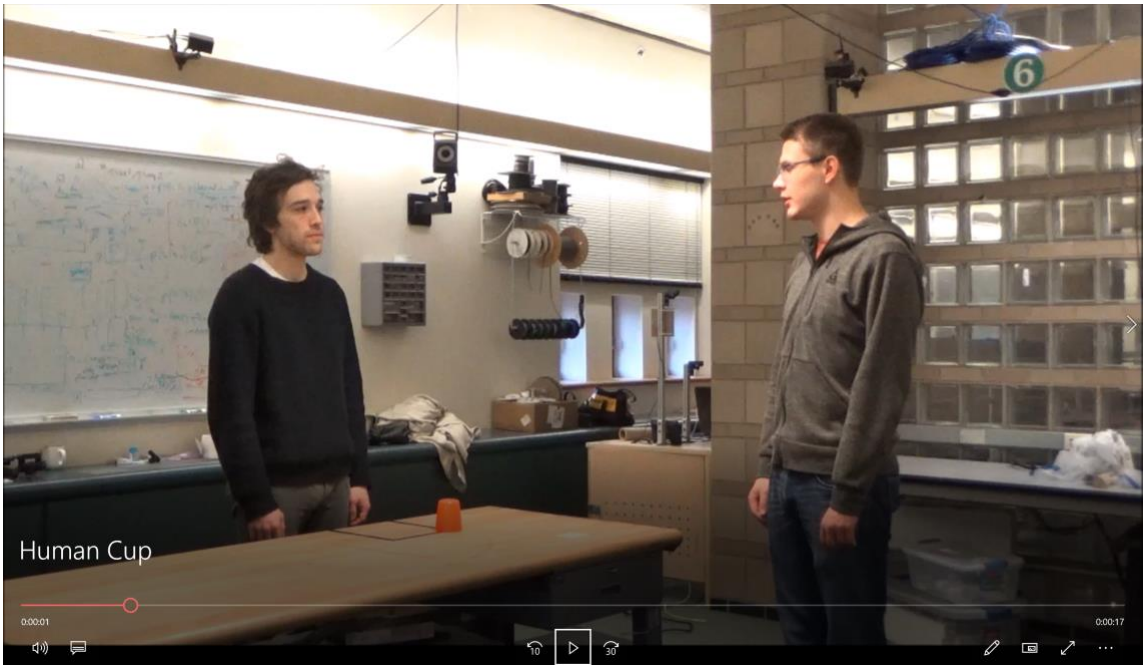


*Figure 1*

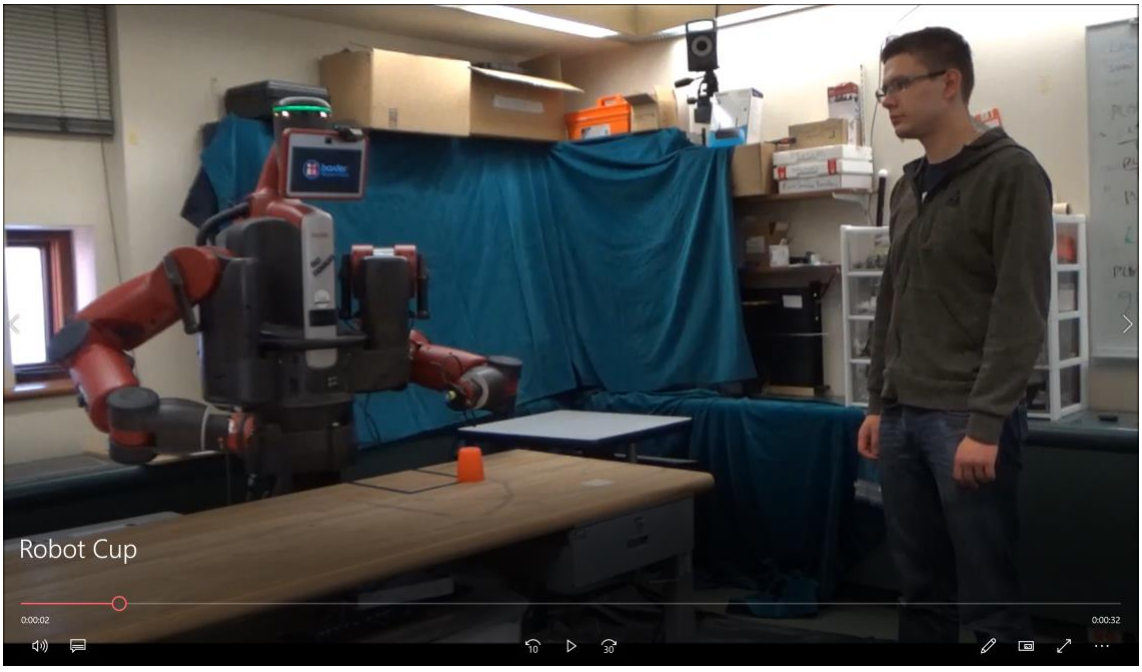


*Figure 2*

On the next page after each video there was a single-line text box asking, "Please enter a short one-sentence summary of what happened in this video, including the consequence." The textbox served as a way to encourage paying attention to the current video and enhancing recall of its contents later. On the next page after the sentence textbox, subjects saw a still photo of the broken object on the table, making the exact extent of the damage clearer than what could be seen during the filmed segment of the actual impact. After the still picture page, participants were asked, "You have just seen a video of damage being done by two agents. Between the commander and the executor, how would you distribute the responsibility for this action? Please enter the percentage of the responsibility (out of 100%) that you would assign to each agent." A hidden timer began at this point, and ended once the participant moved on to the next video, enabling us to measure decision time. Two blank text boxes appeared, each labeled "Commander" or "Executor" in random order with the overhead caption: "Please enter the percentage of the responsibility (out of 100) that you would assign to each role." Both text boxes started by displaying 0%, and a third text box showed the total, which was displayed highlighted in red if it was not exactly equal to 100%. The participant entered a number in each text box to express their judgment of who held more responsibility, and by what fraction, in that particular video.

Each participant watched four such video-question blocks, each containing either a robot or a human executor depending on the assigned condition. The videos were presented in random order, but the actors were consistent; the commander was the same actor for all videos, the robot executor was identical in all robot condition videos, and the human executor was the same actor in all human condition videos. After all four videos were played, we asked participants to rank all the commanders and all the executors against each other. This ranking was accomplished by

asking participants to "Please rank the four commanders you saw on how much they are responsible (where 1 is the most responsibility and 4 is the least)" by placing text bubbles with "The video with a [object]" (where [object] was "cup," "thermos," "mug," or "glass") in vertically ranked order from most to least responsibility within a box labeled "Responsibility Ranking." An identical procedure was completed with the executors. Similarly, participants were asked to rank the videos from most to least damage as a way to check that they paid attention, such that we could exclude from the dataset those who entered an illogical ranking.

At this point, the data collection phase of the study was finished. A debriefing form appeared as a block of text on a new page. In this form, we revealed the nature of the deception we used, and generally explained why we cared about responsibility distribution in the videos. Once the participant navigated to the next page, their participation in the study was complete, and they were given a randomized code to enter into the Amazon Mechanical Turk job such that we could verify that they had completed the study and subsequently deliver our payment to them.

Analysis:

The data we analyzed included participants' gender, age and the intra-and inter-video rankings. Decision times were recorded, but were not analyzed in the present study due to the lack of strong statistical results in other variables to help interpret timing results. 43 of 80 subjects were excluded for failing the manipulation check, which consisted of correctly ranking the four objects on a damage scale, and thus a smaller data set (n = 37) was analyzed. The human versus robot condition assignment balance was unaffected by exclusion, with 18 and 19 subjects, respectively. The average age was 31.22 (SD = 6.07, minimum = 21, maximum = 44) and the subjects were 64.9% male (n = 24, female n = 12, other n = 1).

We examined the effect of damage severity within subjects, as each subject saw the full scale of items regardless of their assigned condition. The independent variable was the object presented, and the dependent variable was the percentage responsibility assigned to the executor (out of 100). To locate this effect and those below, we ran a general linear model (GLM) on the percentages of responsibility assigned to the executor after each of the eight videos (explanatory variables of 2 conditions x 4 videos each). Sphericity was violated (Mauchly's $W = 0.697$, $p = 0.033$), and thus we used the more conservative lower-bound test. The within-subject object effect was not significant in the data ($F = 0.384$, lower-bound $p = 0.539$), failing to support our hypothesis that subjects within one condition would assign different responsibility distributions to the executor and commander based on which object was damaged (Figure 3).

The effect of the human versus robot executor condition was analyzed across subjects, both as a main effect and as an interaction with objects. There was a significant main effect of condition ($F = 7.763$, $p = 0.009$), with human executors blamed more than robot executors (Figure 3). There was no significant interaction effect between condition and object ($F = 0.442$, lower-bound $p = 0.511$), meaning that the condition did not affect how subjects allocated responsibility between objects. Once more, this result fails to support one of our hypotheses.

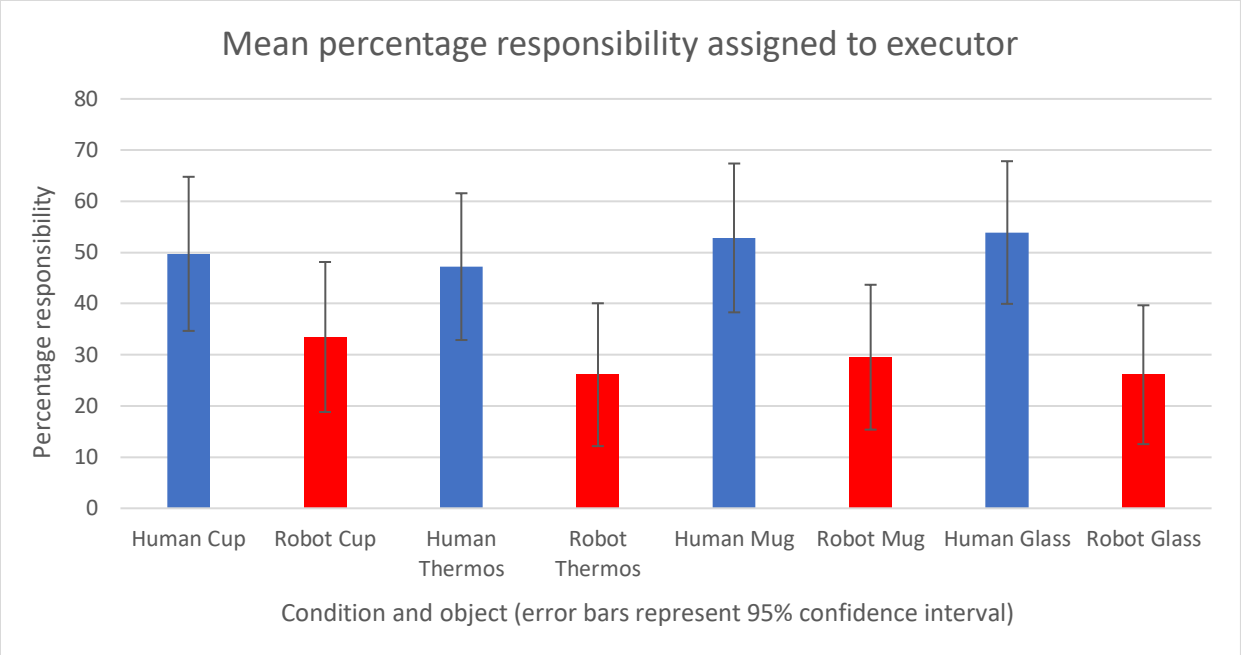


*Figure 3*

Although the percentage of responsibility assigned to the executor showed no significant within-subject effect, the across-video commander responsibility ranking had a significant object effect ($F = 38.369$, lower-bound $p < 0.001$). Sphericity was violated (Mauchly's $W = 0.581$, $p = 0.003$) and thus the conservative lower-bound test was used. As with the responsibility percentage analysis above, we used a GLM with explanatory variables of condition and object, but this time the dependent variable was the responsibility rank (1-4, with 1 being the most responsibility) assigned to the commander from each video. The pairwise comparisons showed significant differences between object ranks (all $p < 0.001$) except for the cup and thermos ($p = 0.106$). There was no effect of condition on commander ranking, but this result is meaningless, as the mean ranking within each condition was by definition 2.5 (on a scale of 1-4). There was also no interaction effect between object and condition ($F = 1.816$, lower-bound $p = 0.186$),

indicating that the assigned condition did not affect how subjects ranked the responsibility of the different commanders they saw (Figure 4).

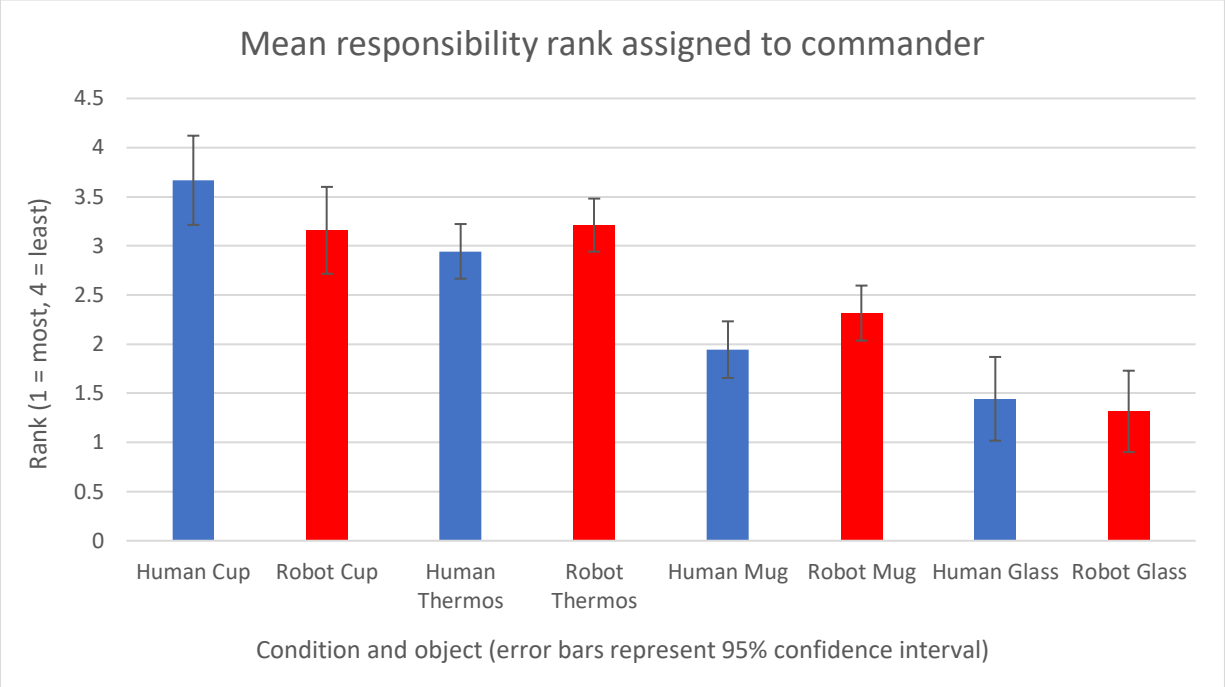


*Figure 4*

The same tests for across-video executor responsibility ranking returned similar results. Sphericity was once more violated (Mauchly's $W = 0.283$, $p < 0.001$). There was a significant effect of object ($F = 6.390$, lower-bound $p = 0.016$). Not all pairwise comparisons showed a significant difference in object ranking. The cup and thermos ($p = 0.154$) and thermos and mug ($p = 0.098$) showed no differences. However, all other pairs had significant ranking differences (all $p < 0.043$). As before, there was no main effect of condition due to the ranking data design. No condition-object interaction effect was present ($F = 0.473$, lower-bound $p = 0.496$), meaning that the condition did not affect ranking order (Figure 5).

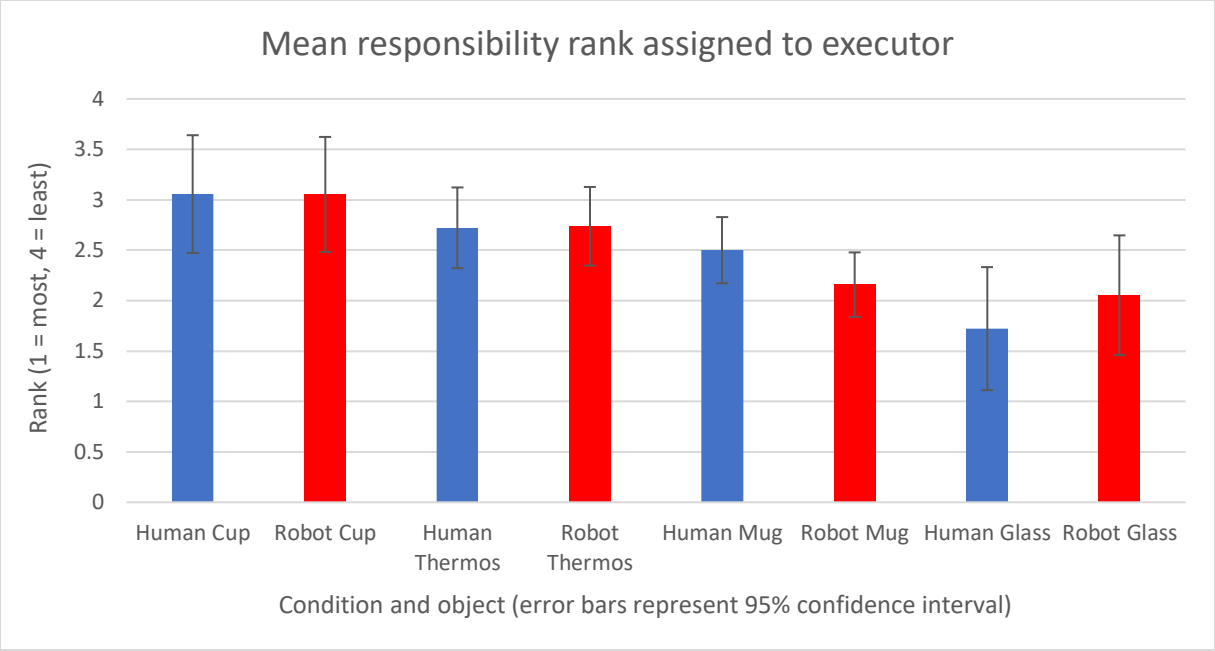


*Figure 5*

Discussion:

We found three strong results in our data. First, those participants who saw the robot executor consistently assigned less responsibility to it than the human condition subjects assigned to their executor (Figure 1). This result may seem unsurprising, given that the robot executor was not human and did not even have a face, and yet all dialogue, actions, and objects were nearly identical between the human and robot conditions. To apply this analysis to the prior example of autonomous cars, there could be two identical crashes, both caused by a human passenger commanding the driver, and a robotic driver would be judged to hold less of the responsibility for the crash than a human driver, holding all else equal. Thus, even though the object and condition-object interaction effects were insignificant, contrary to our hypotheses, the present study provides evidence that people are not willing to assign as much responsibility to a robot as they do to a human, with all other variables held as constant as possible.

The second and third strong results were the significant object effects on the across-video rankings of both commanders and executors. For the two groups, the responsibility ranking correlated with damage level – for example, both roles in the cup videos held lower responsibility ranks than they did in the other videos with more damage. Despite seeming like an important effect, it is likely that subjects were simply replicating the damage rankings, especially when we take into account the fact that the within-subject object effect was insignificant in the executor responsibility percentage analysis. If subjects were allocating responsibility shares out of a whole, then we would expect the commander and executor across-video ranking trends to go in opposite directions; if the commander takes more responsibility as damage goes up, then the executor should take less responsibility with each subsequent step of damage. We, on the other hand, observed that the trends moved in the same direction (Figures 2 and 3), making it unlikely that subjects distributed responsibility dissimilarly across different objects.

As mentioned before, the results of the present study did not support several of our hypotheses, and future work could refine our survey design in order to get clearer results. The largest issue was the high exclusion rate – 43 of our 80 subjects did not agree with our damage ranking, and thus were removed from analysis. The consequences of the videos included an undamaged plastic cup, a badly dented metal thermos, a mug with a broken handle, and a shattered glass, which in our opinion have a clear order in terms of the damage done to them. The fact that so many subjects did not agree with our ordering could imply two possibilities; first, the distinction between our objects was not great enough, and second, subjects were not paying attention to the extent of the damage to the objects, and instead were focused on other aspects of the video. In our effort to use common, recognizable objects, we chose a set that was entirely beverage-related and under $20 in cost. Perhaps this grouping was too close, and further

research could be done with much greater monetary and functional spacing between the breakable objects. However, future studies should take care not to choose objects that have a lot of subjective emotional value on top of their objective monetary value. For the present study, we decided that including objects such as a children's toy, piece of art, or electronic device could introduce numerous confounds involving how much each subject cared about those objects in their own life. For example, a recent parent, an artist, or an avid smartphone user would value those three objects differently than the average person, respectively. While this emotional value is interesting to investigate, it would be almost impossible to control it across subjects with different life experiences and socioeconomic statuses, whereas the objects we picked have low and stable values for everyone.

In terms of subject attention, the prompt we used ("Baxter, can you break it?") may have primed participants to look for a binary outcome of breakage or no breakage, as opposed to a continuum of damage. This theory is supported anecdotally by the descriptive text boxes we collected to promote participants' attention; many remarked upon whether the robot was successful in breaking the object, and very few specified the exact extent of the damage. One caveat, however, is that we asked the participants to write these short descriptions after watching the video, but before seeing the picture of the outcome of the drop seen in the video. Given that the camera view was relatively wide and not particularly high-resolution, it is possible that the exact damage was never clear from the videos alone. A future study could use zoom, larger objects, or a different camera angle to negate the need for a post-damage picture, and thereby make the exact consequences of the executor's actions clear from one viewing of the video.

Another minor issue was the unusual demographics of our subject pool. The United States general population has a median age of 38.1 years and is 49.2% male (CIA, 2018), so

especially given that our study excluded everyone under 18, the subject pool was much younger and more male than the general population. This result is unusual, as prior literature has found that Mechanical Turk workers have very similar demographics to the United States adult population (Huff & Tingley, 2015). A future study could attempt to collect a more representative data set by releasing small Mechanical Turk work batches at predetermined time intervals, thereby negating the effect of temporal skewing of the population, which is exacerbated by the fact that data collection is almost instantaneous (80 subjects in 58 minutes).

Finally, some minor confounding variables exist in our design; a future study could add a face to the robot and perhaps speed up its motion to match that of the human, but otherwise it would be difficult to make the two executors any more similar. We made sure that the human executor did not add any emotion in either his voice or his facial expressions, and maximally slowed his actions, almost to the verge of making them look somewhat unnatural. Unfortunately, the Baxter robot cannot move very quickly, and thus not only did it force the human action to be unusually slow, but it also required the movement combination itself to be very gentle, which contradicted the violent nature of the goal. If the average person were asked to break any container from our set, they would most likely not choose to drop it from a stationary outstretched hand, and the uncanniness of this motion may have affected our results. Additionally, the more artificial a robot looks, the more likely it is that subjects will think that it does not understand that breaking objects is wrong or that commands can be disobeyed, further distancing it from the human executor. In short, our study could be improved in the future by making the two executors more visually similar, both passively and in action.

Conclusion:

Humans are moral creatures, and it is almost unimaginable to view the world of social interactions without assigning moral evaluation to every decision made and action taken. Morals help us to develop a ruleset that keeps society functional and productive, tempering our selfish or animalistic desires into outcomes that are more societally beneficial. As we build machines that can interact with us in our own language, in our most personal and unstructured spaces, they will inevitably enter into the moral system that we have constructed to bring order to the chaos of coordinating immensely intelligent and self-interested beings. While robots lack this latter quality, at least for the time being, they also can make decisions and follow internal goals, even if without free will.

This study sought to examine how naïve participants – those who did not have a professional understanding of how robots work – would distribute the responsibility when a robot was verbally commanded to do an objectively negative action. Would the robot be absolved of all responsibility, or would its human-like voice and form prompt subjects to think of it similarly to the human executor in the other condition? We proposed that as the negative action, which involved breaking household items, got worse, not only would more responsibility fall on the commander relative to the executor, but also that this transfer would be stronger for the robot executors, and that they would be given less responsibility across subjects for the same damage level. After conducting an online survey of Amazon Mechanical Turk users, we found that they did not transfer responsibility in this fashion, failing to support our main hypothesis. However, robot executors were assigned significantly less responsibility than human ones. We hope to see continued work on the subject, especially as more household robots and self-driving cars become available on the market, and the likelihood of blameworthy actions on these robots' part skyrockets. It is of the utmost importance to both commercial and legal interests to see how

the average person will respond to the entrance of intelligent and social machines into their everyday life; we cannot treat robots as humans, but we also cannot ignore them as predictable mechanical appliances. Companies have to design robots that will be pleasant and predictable for customers to interact with, and must prepare for the ways that layperson beliefs will affect jury decisions in court cases related to robots creating undesirable outcomes.

Author Contributions:

Maxim Baranov, Meiying Qin, and Brian Scassellati worked on experimental design, which was initially inspired by redesigning an unpublished study by Qin. Baranov wrote the IRB proposal with edits by Qin. Baranov wrote and edited the first draft of the paper, borrowing some background from the aforementioned unpublished work by Qin. Baranov and Qin collected data. Qin and Baranov conducted data analysis, and Baranov wrote the final draft with input from Qin and Scassellati.

Works Cited:


CIA. (2018, April 02). The World Factbook: United States. Retrieved from https://www.cia.gov/library/publications/the-world-factbook/geos/us.html

Dale, R. (2016). The return of the chatbots. *Natural Language Engineering*, *22*(5), 811-817.

DeBaets, A. M. (2014). Can a robot pursue the good? Exploring artificial moral agency. Journal of Evolution and Technology, 24, 76-86.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, *14*(3), 349-379.

Hartley, R., & Zisserman, A. (2003). *Multiple View Geometry in Computer Vision*. Cambridge University Press, xi.

Hinds, P. J., Roberts, T. L., & Jones, H. (2004). Whose job is it anyway? A study of human-robot interaction in a collaborative task. Human-Computer Interaction, 19(1), 151-181.

Huff, C., & Tingley, D. (2015). "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research & Politics*, *2*(3), 2053168015604648.

Lee, S. L., & Lau, I. Y. M. (2011, March). Hitting a robot vs. hitting a human: is it the same?. In *Proceedings of the 6th international conference on Human-robot interaction* (pp. 187-188). ACM.

Kahn Jr, P. H., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., ... & Severson, R. L. (2012, March). Do people hold a humanoid robot morally accountable for the harm it causes?. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction* (pp. 33-40). ACM.

Knight, W. (2015, September 16). The Roomba Now Sees and Maps a Home. *MIT Technology Review*. Retrieved from https://www.technologyreview.com/s/541326/the-roomba-now-sees-and-maps-a-home/

Kozub, S. (2017, March 15). You can now use Alexa to control your Roomba. *The Verge*. Retrieved from https://www.theverge.com/circuitbreaker/2017/3/15/14933636/alexa-roomba-voice-commands-irobot-home-app

Kwak, S. S., Kim, Y., Kim, E., Shin, C., & Cho, K. (2013, August). What makes people empathize with an emotional robot?: The impact of agency and physical embodiment on human empathy for a robot. In *RO-MAN, 2013 IEEE* (pp. 180-185). IEEE.

Malle, B. F., Scheutz, M., Forlizzi, J., & Voiklis, J. (2016, March). Which Robot Am I Thinking About?: The Impact of Action and Appearance on People's Evaluations of a Moral Robot. In The Eleventh ACM/IEEE International Conference on Human Robot Interaction (pp. 125-132). IEEE Press.

Naturalsoft Ltd. (2018). Free Text to Speech Online with Natural Voices. Retrieved from https://www.naturalreaders.com/online/

Prassler, E., Ritter, A., Schaeffer, C., & Fiorini, P. (2000). A short history of cleaning robots. *Autonomous Robots*, *9*(3), 211-226.

Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., & Taylor, S. H. (2017, May). Alexa is my new BFF: social roles, user satisfaction, and personification of the amazon echo. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 2853-2859). ACM.

Rethink Robotics. (2018). Baxter Collaborative Robots for Industrial Automation. Retrieved from http://www.rethinkrobotics.com/baxter/

Sparrow, R. (2007). Killer robots. Journal of applied philosophy, 24(1), 62-77.

Sullins, J. P. (2006). When Is a Robot a Moral Agent?. International Review of Information Ethics, 6, 12.

Van Camp, J. (2017, November 7). Review: Jibo Social Robot. *Wired*. Retrieved from https://www.wired.com/2017/11/review-jibo-social-robot/