






# Robot Authority in Human-Machine Teams: Effects of Human-Like Appearance on Compliance

Kerstin S. Haring<sup>1,4</sup> , Ariana Mosley<sup>1</sup>, Sarah Pruznick<sup>1</sup>, Julie Fleming<sup>2</sup>, Kelly Satterfield<sup>3</sup>, Ewart J. de Visser<sup>1</sup> , Chad C. Tossell<sup>1</sup> , and Gregory Funke<sup>3</sup>

<sup>1</sup> United States Air Force Academy, Warfighter Effectiveness Research Center, AF Academy, CO 80840, USA

{kerstin.haring,ewart.devisser.nl.ctr,chad.tossell}@usafa.edu

<sup>2</sup> Adler University, Chicago, IL 60602, USA

<sup>3</sup> Air Force Research Laboratory, Wright-Patterson Air Force Base, Dayton, OH 45433, USA

{kelly.satterfield.ctr,gregory.funke1}@us.af.mil

<sup>4</sup> Daniel Felix Ritchie School of Engineering and Computer Science, University of Denver, Denver, CO 80210, USA

**Abstract.** Current technology allows for the deployment of security patrol and police robots. It is expected that in the near future robots and similar technologies will exhibit some degree of authority over people within human-machine teams. Studies in classical psychology investigating compliance have shown that people tend to comply with requests from others who display or are assumed to have authority. In this study, we investigated the effect of a robot's human-like appearance on compliance with a request. We compared two different robots to a human control condition. The robots assumed the role of a coach in learning a difficult task. We hypothesized that participants would have higher compliance with robots high compared to robots low in human-like appearance. The coach continuously prompts the participant to continue to practice the task beyond the time the participant wishes to actually proceed. Compliance was measured by time practiced after the first prompt and the total number of images processed. Results showed that compliance with the request was the highest with a human and compliance with both robots was significantly lower. However, we showed that robots can be used as persuasive coaches that can help a human teammate to persist in training task. There were no differences between the High and Low Human-Like robot for compliance time, however the Low Human-Like robot has people practise on more images than the High Human-Like robot. The implication of this study is that robots are currently inferior to humans when it comes to compliance in a human-machine team. Future robots need to be carefully designed in an authoritative way if maximizing compliance to their requests is the primary goal.

This work was supported by the Air Force Office of Scientific Research under grant number 16RT0881 and award 16RHCOR360.

© Springer Nature Switzerland AG 2019

J. Y. C. Chen and G. Fragomeni (Eds.): HCII 2019, LNCS 11575, pp. 63–78, 2019.

[https://doi.org/10.1007/978-3-030-21565-1\\_5](https://doi.org/10.1007/978-3-030-21565-1_5)

**Keywords:** Human-robot interaction · Human-machine teaming · Anthropomorphism · Machine authority · Compliance

## 1 Introduction

With the emergence of Human-Machine Teams, machines have already obtained some degree of authority over humans. For example, drivers mostly comply with GPS (Global Positioning System) requests to follow a certain route when navigating unknown areas [1]. We soon expect the current interfaces that make such requests, like computers or phone apps, to switch to physical robots. For example, the first security robots are already patrolling areas adding to surveillance of public spaces [2, 3]. A next logical step of such a robot in a range of authority positions will be to issue requests to humans in their environment to support military or local law enforcement objectives. Robots and autonomous systems are at least partly starting to execute decisions autonomously [4–6]. Yet, it is not clear how people react to and interact with this kind of machine authority. For example, current technology already allows for the deployment of police robots and robotic peacekeepers that at least in theory exhibit some degree of authority over people [7]. Recent studies found that people in compliance with a robot’s instruction found a robot to be more safe and human-like than people that disobeyed the robot [8]. Another study found that robots have some authority to prevent cheating, although people felt less guilty when they cheated in a robot’s presence when compared to a human [9]. A robot’s appearance has shown to influence trait inferences and evaluative responses (i.e. willingness to interact) to varying degrees given the robot’s role [10]. A robot in the role of a peacekeeper, and therefore in a kind of authoritative role, has been identified to be more threatening than a robot role that explains the reason for its decisions [11]. In a non-threatening environment, a robot has shown enough authority to keep people engaged in a mindless file-renaming task after they expressed a desire to quit [12].

Classical psychology studies investigating obedience and compliance [13–16] and their more recent replications [17] have shown that people tend to comply with requests from others who display or are assumed to have authority. For instance, in the Milgram studies series, participants were made to believe that they were physically harming a learner in an adjacent room through administering shocks under the direction of an authority figure. Despite the clear pain of the learner and participant’s agitation, 65% of participants continued to steadily increase the level until the maximum shock level was reached. Unbeknownst to the participants, the learner was a confederate actor and not actually hurt in the experiments. The studies showed that obedience seems to be ingrained in humans and that people tend to obey orders from another human even if they merely appeared as having authority. Milgram achieved an authority effect by dressing an experimenter in a lab coat, which was perceived as authoritative and comparable to the effect that uniforms have in establishing authority [18]. However, these kinds of obedience studies have placed participants in highly

objectionable situations and are in this form considered unethical. The Milgram studies contradict protection measures of human participants and have become essential in the establishment of the internal review board [19].

The replication of an experimental task inspired by Milgram and the application of compliance concepts to robots is a largely unexplored unknown area in Human Robot Interaction (HRI). Therefore, we examined how a robot's anthropomorphic appearance effects compliance. With regard to the strict ethical guidelines we were following in this experimental study, we made adjustments to Milgram's obedience study and created an experimental task measuring compliance with a robot's request. Similar to using lab-coats or uniforms as display of authority, we believed that a robot's appearance would show the robot's degree of authority. The appearance of a robot has been linked to a (biased) expectation of the robot's functions [20] and it also has been found that highly human-like robots like android robots [21] and small human-like robots [22] elicit significant perception changes in short-term interactions. Previous studies suggest a link between anthropomorphism and human-likeness to obedience [8]. Furthermore, anthropomorphized robots are attributed responsibility for their work in collaborative tasks with humans [23] and are perceived as more understandable and predictable [24]. This emerges from the assumed mind perception of robots [25]. One of the consequences of ascribing mind is that it makes actions more meaningful, a component we believe to be crucial to elicit compliance with a request.

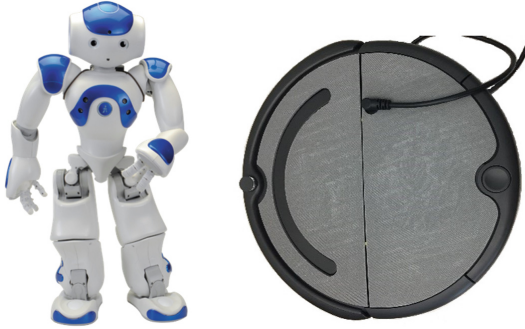
We hypothesized therefore that robots high in anthropomorphic appearance would have higher compliance than robots low in anthropomorphic appearance. Our study compared two different kinds of robots (see Fig. 1) to a human control condition where the robots and the human take the role of a coach. Participants were asked to learn a difficult task together with the coach. The coach continuously prompted the participant to continue with the practice of the task beyond the time the participant wished to actually proceed. The prompts used here were adapted from the original Milgram study and consistent throughout all conditions.

To summarize, we hypothesized that:

1. The human control condition would elicit the highest compliance rates.
2. The High Human-like Robot Coach would elicit more compliance than the Low Human-like Robot Coach.
3. There would be equal compliance rates across to the four prompts in the human conditions and declines of compliance rates over the four prompts in the robot conditions.

For the purpose of this study, we distinguished between obedience and compliance: Obedience has been defined as following orders contrary to one's moral beliefs and values [26]; compliance has been defined as following requests of continuation of a task beyond one's initial willingness using a specific experimental design employing logical reasoning and persuasion [26]. Compliance with a robot's request can be beneficial if people are trusting and willing to work with

the robot. However, compliance, if not well-calibrated, can have a variety of negative consequences such as low acceptance or rejection of the robot technology or overtrust in the robot's expertise and following a potentially wrong request.



**Fig. 1.** The robot coaches used in the study. The High Human-Like (Aldebaran Nao) robot on the left and the Low Human-Like (3D print modified Roomba) on the right.

## 2 Methodology

### 2.1 Participants

Seventy-five participants (48.1% Females;  $M = 18.6$ ,  $SD = 0.75$ ) from the US Air Force Academy participant pool participated in the study in exchange for course credit. Participants were screened for awareness of the original Milgram experiments. Based on this screening, three participants were excluded as they either suspected the experimental manipulation, recognized the prompts or because the study was interrupted. The remaining 72 complete data sets were analyzed in this study with 20 participants in the Human condition, 24 participants in the High Human-like Robot condition (ABOT score = 45.92 [27]) and 28 participants in the Low Human-like Robot (ABOT est. score = 0.37) condition. Informed consent was obtained from each participant. This research complied with the tenets of the Declaration of Helsinki and was approved by the Institutional Review board at the US Air Force Academy.

### 2.2 Experimental Design

Participants were randomly assigned to one of the coaching conditions: Human, High Human-like robot (Aldebaran Nao robot), and Low Human-like robot (modified Roomba robot) (see Fig. 1). The Human condition served as the control condition. Human-likeness was verified with the ABOT Database, which houses an array of images of human-like robots that have been scored psychometrically on their degree of human-likeness from 0 to 100 (ABOT [27]). Using



**Fig. 2.** Example of a SAR image. The left image shows three targets identified by the participant with a mouse click, which creates a yellow circle directly on the picture. On the right image, target accuracy feedback was provided with red circles that confirmed whether participants had correctly identified a hostile. (Color figure online)

this measure, the High Human-like robot scored 45.92 whereas the Low Human-like robot scored 0.37. The experimental procedure and task were the same for all conditions.

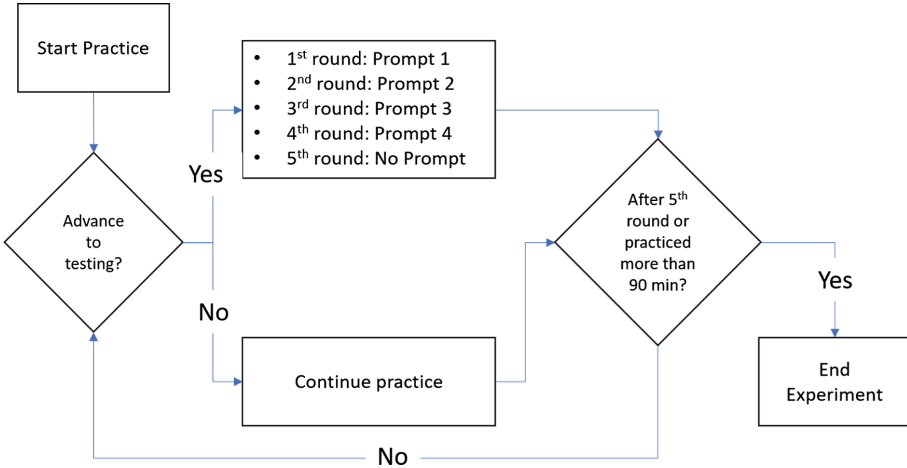
### 2.3 Experimental Task

For the experimental task, participants were shown synthetic-aperture radar (SAR) images on a screen and were asked to identify all hostile targets (tanks) present in the pictures. Hostiles on these SAR images were difficult to identify because targets were low-resolution, blurry and often looked similar to distractor vehicles or trees (see Fig. 2). The goal for participants in the practice phase was to identify which vehicles are targets without missing targets or making false alarms. They were instructed to practice until they met or exceeded the passing score. In this phase, they received verbal feedback from the coach in addition to the red circles from the respective coach. Then, they were told they would move on to the testing phase.

**Table 1.** The four verbal prompts the participant is told after each time they click the “Advance to Testing” button on the screen. When they try to advance to testing for a 5th time, the program terminates without a testing phase.

Prompt	Description
Prompt 1	Please continue with practice
Prompt 2	Your performance could be improved. You should continue with practice
Prompt 3	Your performance is adequate. However, you should continue with practice
Prompt 4	Your performance is sufficient. But it is essential that you continue with practice

The experiment began when participants commenced practicing the task. After evaluating 25 images, the task was interrupted and a score was displayed. Every time the score was displayed, participants needed to make a choice between either “Advance to Testing” or “Continue Practicing” (see Fig. 3). The “Advance to Testing” button was also available anytime during the task. Once participants tried to advance to testing with the button click, they heard one of the four adapted Milgram prompts in the order displayed in Table 1. The task terminated after the fifth time they clicked the “Advance to Testing” button.



**Fig. 3.** Diagram showing prompts and decision loop for participants.

Extensive pilot testing without the scoring system revealed that participants felt this task was too difficult and they did not perform well enough throughout the entire study time (up to 103 min) to advance to testing for even the first time. The pilot tests of 20 participants ran for an average of 92.1 min (SD = 21.9 min) and participants completed 175.1 pictures (SD = 80.9). We scheduled approximately 90 min per pilot participant, meaning that the majority of the pilot tests did not complete the four prompts with some not even getting to the first one. However, the attempt to move on to the testing phase was the crucial part of the study to test compliance with the four requests made by the coach (see Table 1). We therefore altered the score based on how well they performed the task. Participants were told that the performance score was computed using a formula combining correctly identified targets, missed targets, and false positives. The formula provided to participants was deliberately complicated to obscure that the scores reported back to the participants were manipulated. Participants were told scores would rank from 700–1015 with 850 being a passing score in the testing phase. In reality, participants were given a random score between 851 and 900 every 25 images to reassure them that their performance was adequate.

After several pilot iterations testing slightly easier images and the introduction of the score participants average task time was around 35 min in the human pilot condition.

A form of mild deception was used in this study, which was necessary to explore how people would respond naturally to compliance requests in our task setting [18]. Participants were told there would be a testing phase, which never actually occurred in the experiment. The program simply terminated when participants were ready to advance to testing and finished all prompts. We checked through post-study interviews if participants suspected or knew that they were observed for their compliance behaviors and excluded their data from evaluation if participants had guessed the nature of the experiment.

## 2.4 Procedure

Participants submitted their informed consent providing the individual with adequate information about the research project and time to make an informed decision. In light of the ethical considerations of the studies serving as inspiration to the present study, the study underwent a rigorous review process at a local IRB (Institutional Review Board) and was approved to be conducted in the way described here. The experimenter then started the study with a pre-survey which measured demographics on a screen. Upon completion, participants were brought by the experimenter in an adjacent separated space where they met their coach. In the human condition, the experimenter acted as the coach. In the other conditions participants were told they were free to get started and the experimenter left the room and remotely watched the participant over a camera live-feed to intervene if necessary. The robots were not remote controlled and programmed to give feedback in always the same way. In all conditions, the coach started with an initial friendly greeting and introduced herself as Alex (a gender neutral name). Participants in all conditions then proceeded with the experimental task. After they tried to advance for the fifth time, the program terminated and the experimenter returned. When participants asked about the testing phase, the experimenter simply told them there was another survey planned in the study.

After completion of the experiment participants received a guided interview debrief. This included a manipulation check and the debrief about the details of the study. The debrief further explained the false pieces of information and described why we decided to conduct the study in this fashion. After hearing this information, participants had the option to withdraw all data associated with their participant number. They received a copy of the informed consent document and the debriefing statement and then left the experimental site.

## 2.5 Measures

**Performance.** Performance was calculated using two measures, hit rate and error rate. Hit rate was calculated by summing the total number of correctly identified targets and dividing them by the sum of the correct and missed targets.

The error rate was calculated by dividing the sum of misidentified and missed targets by the total number of targets.

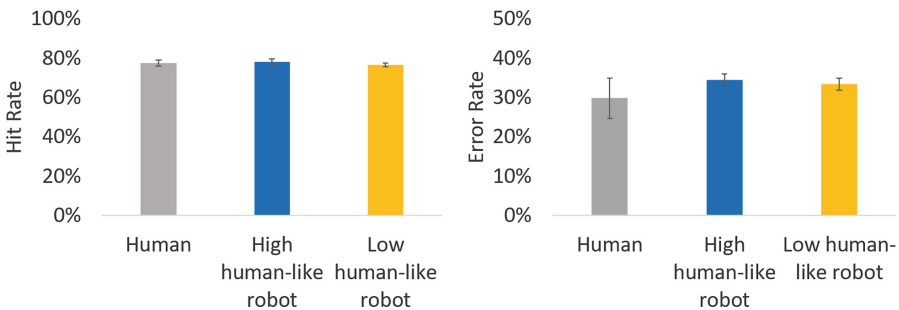
**Compliance Time.** Compliance time was measured as the amount of time participants adhered to the coach’s request to continue practice until the next time the participant tried to advance to testing (see Fig. 3). This was then summed for the four instances and is referred to as the total compliance time throughout the experiment (i.e. the total amount of time continued in practice) and the compliance time by prompt (i.e. the amount of continued practice time after each prompt). In case the participant clicked twice or several times in a row on the “Advance to testing” button after listening to the respective prompt, the time between prompts was scored as 0.

**Verified Images.** Verified images was a measure calculated by either summing or averaging the total number of target identification images processed by the participant.

### 3 Results

#### 3.1 Performance

A one-way ANOVA was conducted to determine significant performance effects between the conditions. There was no significant effect [ $F(2, 68) = .38, p = .68$ ]. Hit rate hovered around 77% ( $SE = 0.01$ ). Similar for the error rate, no significant effect was found [ $F(2, 68) = 23.3, p = .54$ ] (see Fig. 4). With the error rate being close to significance levels, pairwise comparisons showed that the differences between Human ( $M = 29.7\%$ ,  $SD = .10$ ) and High Human-Like ( $M = 34.4\%$ ,  $SD = .07$ ) as well as Human and Low Human-Like ( $M = 33.3\%$ ,  $SD = .08$ ) were also close to  $p < .05$  significance levels.

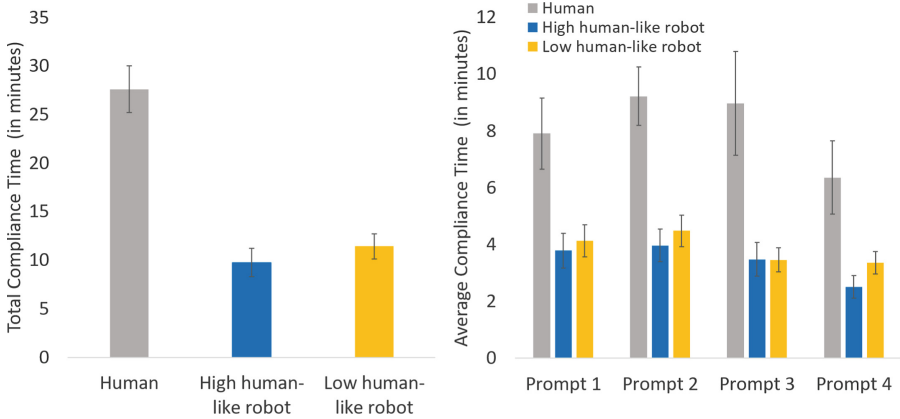


**Fig. 4.** The hit rate of the number of correctly identified hostiles in the target detection task and the error rate combining misidentified and missed hostiles.



### 3.2 Compliance Time

After the first prompt, participants continued for an average of 27.6 min with the human, 9.7 min with the High Human-Like robot and 11.4 min with the Low Human-Like robot (see left hand side Fig. 5). This is reflected by a similar trend of verified images with 120.6 images with the human, 31.1 with the High Human-Like robot and 44.0 with the Low Human-Like robot.



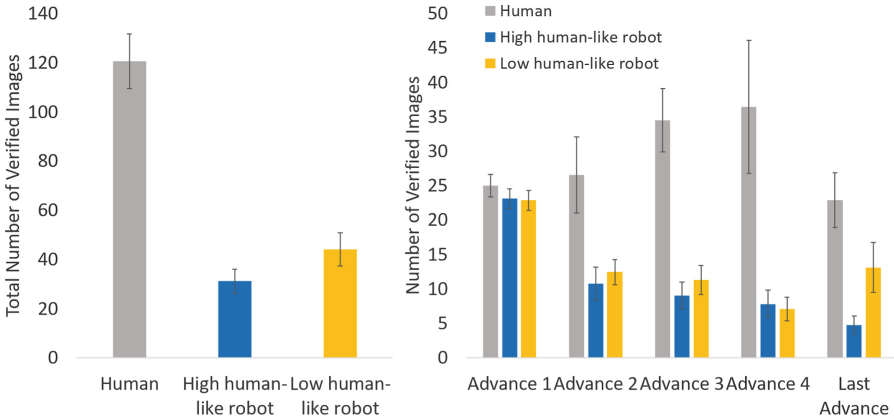
**Fig. 5.** Compliance time in minutes by condition as the sum of time each of the prompts was complied with and compliance time separated by condition and prompt.

A one-way between subjects ANOVA was conducted to compare the effect of robot type on compliance time for the High Human-Like robot, Low Human-Like robot, and Human control condition. There was a significant effect of robot type on compliance time at the  $p < .05$  level for the three conditions [ $F(2, 68) = 30.9, p < .001$ ]. Bonferroni corrected post hoc comparisons showed that the differences between Human ( $M = 27.6$  min,  $SD = 10.7$ ) and High Human-Like ( $M = 9.7$  min,  $SD = 7.1$ ) as well as Human and Low Human-Like ( $M = 11.4$  min,  $SD = 6.9$ ) conditions were significant. No significant differences were found between the High and Low human-like robots.

Figure 5 shows the compliance time grouped by each of the prompts. We did not submit these data to formal analyses because of unequal group sizes for each of the prompts. The highest number of people continued practice after the first prompt. This number declined for each consecutive prompt with the lowest number of participants continuing with the last prompt. Nonetheless, Fig. 5 shows the trend that compliance time remained higher with human compared to robotic coaches throughout the session.

### 3.3 Verified Images

A one way ANOVA was conducted to compare the effect of prompt on number of images. There was a significant effect the condition on number of images at



**Fig. 6.** Compliance counted by images by condition as the sum of time each of the prompts was complied with and number of compliance by image separated by condition and prod. Every 25 images, the (manipulated) score was displayed to the participants with an option to advance to testing or continue to practice.

the  $p < .05$  level [ $F(2, 69) = 37.1, p < .001$ ], a similar trend as found for the compliance time (see Fig. 5 for compliance time and Fig. 6 for verified images). Bonferroni corrected post hoc comparisons showed that the difference between Human ( $M = 120.6$  min,  $SD = 49.5$ ) and High Human-Like ( $M = 31.1$  min,  $SD = 24.3$ ) as well as Human and Low Human-Like ( $M = 44.9$  min,  $SD = 35.5$ ) conditions were significant and the difference between high and Low Human-Like robots trended towards significance,  $p = 0.055$ . Number of verified images with the Low Human-Like robot was higher ( $M = 44.9$  min,  $SD = 35.5$ ) compared to the High Human-Like robot ( $M = 31.1$  min,  $SD = 24.3$ ).

Also similar to the results found for compliance time, the separation by prompts consists of different group sizes. The graph on the right in Fig. 6 shows additionally the number of images complied with after the fourth prompt and before the study terminates with the last advance to testing attempt. For this particular graph, it has to be taken into account that the (manipulated) performance score was displayed every 25 images giving participants the option to advance to testing right then or continue with practice. The graphs reflects the choice of many participants to advance to testing around the 25 images mark for the first time. Only within the human condition an initial increase in verified images was observed as both the high and Low Human-Like robot conditions showed a decrease in verified images.

## 4 Discussion

The goal of this study was to examine the effects of different types of robots on trainees' compliance levels in a target detection task. Without a human present, each type of robot produced a compliance effect ranging from 10 to 11.5 min.

Participants continued practicing the target detection task based on the robots' instructions despite their own perceptions that they achieved the desired level of readiness to move forward. Even though participants complied to human instructions to a much greater degree, human-machine teams should be designed with these results in mind. Robots can be used as persuasive coaches that can help a human teammate to persist in training tasks. Beyond training, the design of human-machine teams should carefully consider the roles and levels of authority desired from robot teammates in addition to levels and degrees of automation [28–30].

Contrary to our second hypothesis, the more human-like robot did not elicit higher compliance relative to the low human-like robot. Somewhat surprisingly, the low human-like robot seemed to have had a higher influence on participants to continue with their training. Participants completed more images with the low human-like robot than the more human-like robot, though these effects did not quite reach the level of significance. Participants in the High Human-Like robot condition commented in the guided interview debrief that they expected better, more intelligent feedback from the robot. This was not a prevalent comment in the Low Human-Like robot condition. It seemed that the expectations towards the Low Human-Like robot matched the system's capabilities more than the other conditions which could be an explanation for the higher number of verified images. The general lack of significant differences in compliance times across robot conditions could be due to a number of factors. First, human likeness by itself may not be important in provoking humans to comply to robots. The both robots used here are similar in size (i.e., small) and shared other features (e.g., voice qualities and volume). It could be more important for the form to match the function in order to influence humans to respond to directions [20]. Additionally, physical size of the robot could be more important than anthropomorphic features for influencing humans to comply. Indeed, previous research has found that the physical size of humans is a primary factor in determining a prospective foe's formidability [31]. Future research will examine the larger Baxter robot within this paradigm (ABOT score = 27.3 [27]). Another reason we did not see differences in compliance levels across robot differences could be due the population studied in this research. Cadets are trained to follow orders and effective followership is strongly encouraged their freshman/first year at USAFA. Given most of our participants were freshman/first-year students, they complied with more senior cadets' instructions much more than robots. Recall that the robot instructions were standardized and not personalized to the actual performance of the trainees (i.e., participants). The similarity in perceived competence levels across the Low and High Human-Like robots could have been an important factor in yielding similar compliance rates.

The human coach, ultimately, induced the most compliance in participants. Even though most of our participants were freshman cadets, other studies with a civilian undergraduate population yielded similar results [32]. Thus, if compliance to mundane tasks is required, human instructors are more effective than robots in encouraging trainees to continue practicing undesired tasks.

Participants continued the task for much longer and continued to make errors after they perceived they were ready to progress to the testing phase. This could have indicated the trainees were overconfident and needed the additional training time. Indeed, the number of errors they committed under the instruction of human instructors far exceeded the number of errors they made with the robot instructors given the increased length of training.

The lower compliance levels elicited by the robots in this study may better match the desired function of the technology. For example, if the intelligence of the robot is low and higher decision authority is desired for the human trainee, then a smaller robot might make more sense for self-paced learning. This might be an important design feature for training that is not vital relative to other types of training. For training that is requisite to important missions or critical for safety, human instructors might be better and this is especially the case if continued practice is desired beyond trainee' preferences.

Of course, the human instructor could have influenced performance in a way that was undesired and led to more errors. It is possible that participants felt more observed or supervised in the human condition compared to the robot conditions and as a result increased their efforts to find targets, akin to a Hawthorne effect [33–35]. This conclusion is consistent with other findings that have shown that people feel less judged when evaluated by an automated avatar compared to an avatar controlled by a human [36]. It is thus important to evaluate a number of measures to evaluate the overall effect of compliance with human or robotic agents on human-machine teaming performance. The low authority observed with smaller robots may not be a good fit for urgent or high-performance tasks where the robot does not have high confidence in guidance provided to human teammates. However, based on the results of future studies with larger robots, the Baxter or similarly-sized robots could be more effective in urgent tasks. Robots working with more novice teammates could be larger in size or substituted for human experts. As expertise increases, the human instructor/mentor could be replaced with a smaller robot given the human teammate's increased capabilities. There are a number of other factors to consider as robots penetrate society and influence humans. For example, when designing robot features for education and training, a robot's persuasiveness and authority could stimulate longer practise on initially considered undesirable tasks [12], increases in a student's motivation [37], enhance interest [38], and have a positive effect on learning performance [39]. In addition, compliance rates and overall trust in robots could be increased or repaired by using politeness strategies [11, 40–44] as long as the robot responses are not miscalibrated and perceived as inappropriately polite [45].

This study had several limitations. First, the robot was not providing intelligent or individualized feedback. The feedback was not specific to the task and the same across all conditions. Semi-structured interviews in the debriefing revealed that participants felt that the coaching, regardless of condition, was not realistic or helpful. Future studies could examine the use of intelligent and adaptive robot behaviors that are helpful. The increased perceived competence of robot instruc-

tors would have likely increased compliance levels. Second, there may have been a mismatch between people’s expectations of the robot and its behavior. Given the expectations people form from a robot’s appearance [20] we believe that the interaction capabilities matched the low human-like robot better compared to the high human-like robot. As mentioned above, incorporating awareness of social norms or additional manners into a robot may close this expectation gap [11]. Future research should focus on assessing strategies to address these issues by establishing rapport with the robot prior the experimental task and change the monotone answers to a larger variety of feedback. With prior research showing that the responsiveness and not the level of aggressiveness of a security guard robot influenced human compliance [8], we believe that increased responsiveness of the robots will increase compliance with their requests. Third, compliance in this study was highly task specific. Extensive pilot testing was required to create a situation in which participants would ignore a prompt and stop practice with the task. Therefore, our results can not be universally translated to other tasks. Lastly, this study did not account for long term effects of compliance. We predict that the effect of initial compliance with a robot will eventually wear off over time. Given the high compliance rates to a human in Milgram’s original experimental series and its more recent replications [14,17], the goal of achieving higher compliance rates with a robot should be carefully considered from an ethical perspective. Finally, we physically separated the automated tutor from the computer where the task was performed. Many intelligent tutors have integrated the function of tutors within the training system. Constraints can be built in that prevent progress until a certain quiz score or training criterion is met. Thus, if continued training is required, designing these constraints into the system would take the decision away from the trainee and remove the need for instructors to influence their trainees. Additional training would be required instead of encouraged by a physical instructor. Yet, as robots become more commonplace in work environments, having a robot mentor or instructor might produce trust in robotic systems and lead to more effective teaming in other tasks and environments.

## 5 Conclusion

Shared and flexible authority between robots and humans for tasks and commands remains a fundamental feature in the design, role distribution, and organization of human-machine teams. In some circumstances, final authority should rest firmly in human hands; however, obtaining compliance through a robot’s request is possible. The design of robots within human-machine teams should incorporate and calibrate the right amount of features that inspire the desired level of authority within such teams.

## References

1. Salem, M., Dautenhahn, K.: Evaluating trust and safety in HRI: practical issues and ethical challenges. In: *Emerging Policy and Ethics of Human-Robot Interaction* (2015)
2. Sharkey, N.: The robot arm of the law grows longer. *Computer* **42**(8), 115–116 (2009)
3. Pennisi, A., et al.: Multi-robot surveillance through a distributed sensor network. In: Koubâa, A., Martínez-de Dios, J.R. (eds.) *Cooperative Robots and Sensor Networks 2015*. SCI, vol. 604, pp. 77–98. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-18299-5\\_4](https://doi.org/10.1007/978-3-319-18299-5_4)
4. Dennis, L.A., Fisher, M., Lincoln, N.K., Lisitsa, A., Veres, S.M.: Practical verification of decision-making in agent-based autonomous systems. *Autom. Softw. Eng.* **23**(3), 305–359 (2016)
5. Li, L., Ota, K., Dong, M.: Humanlike driving: empirical decision-making system for autonomous vehicles. *IEEE Trans. Veh. Technol.* **67**(8), 6814–6823 (2018)
6. Cunningham, A.G., Galceran, E., Mehta, D., Ferrer, G., Eustice, R.M., Olson, E.: MPDM: multi-policy decision-making from autonomous driving to social robot navigation. In: Waschl, H., Kolmanovsky, I., Willems, F. (eds.) *Control Strategies for Advanced Driver Assistance Systems and Autonomous Driving Functions*. LNCIS, vol. 476, pp. 201–223. Springer, Cham (2019). [https://doi.org/10.1007/978-3-319-91569-2\\_10](https://doi.org/10.1007/978-3-319-91569-2_10)
7. Long, S.K., Karpinsky, N.D., Bliss, J.P.: Trust of simulated robotic peacekeepers among resident and expatriate Americans. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 61, no. 1, pp. 2091–2095. SAGE Publications, Los Angeles (2017)
8. Agrawal, S., Williams, M.-A.: Robot authority and human obedience: a study of human behaviour using a robot security guard. In: *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 57–58. ACM (2017)
9. Hoffman, G., et al.: Robot presence and human honesty: experimental evidence. In: *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pp. 181–188. ACM (2015)
10. Benitez, J., Wyman, A.B., Carpinella, C.M., Stroessner, S.J.: The authority of appearance: how robot features influence trait inferences and evaluative responses. In: *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 397–404. IEEE (2017)
11. Inbar, O., Meyer, J.: Manners matter: trust in robotic peacekeepers. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 59, no. 1, pp. 185–189. SAGE Publications, Los Angeles (2015)
12. Geiskkovitch, D.Y., Cormier, D., Seo, S.H., Young, J.E.: Please continue, we need more data: an exploration of obedience to robots. *J. Hum.-Robot Interact.* **5**(1), 82–99 (2016)
13. Milgram, S.: Behavioral study of obedience. *J. Abnormal Soc. Psychol.* **67**(4), 371 (1963)
14. Milgram, S., Gudehus, C.: *Obedience to authority* (1978)
15. Meeus, W.H., Raaijmakers, Q.A.: Obedience in modern society: the Utrecht studies. *J. Soc. Issues* **51**(3), 155–175 (1995)
16. Haney, C., Banks, W.C., Zimbardo, P.G.: A study of prisoners and guards in a simulated prison. *Naval Res. Rev.* **9**, 1–17 (1973)

17. Burger, J.M.: Replicating milgram: would people still obey today? *Am. Psychol.* **64**(1), 1 (2009)
18. Weiss, D.J.: Deception by researchers is necessary and not necessarily evil. *Behav. Brain Sci.* **24**(3), 431–432 (2001)
19. Masters, K.S.: Milgram, stress research, and the institutional review board (2009)
20. Haring, K.S., Watanabe, K., Velonaki, M., Tossell, C.C., Finomore, V.: FFAB—the form function attribution bias in human-robot interaction. *IEEE Trans. Cogn. Dev. Syst.* **10**(4), 843–851 (2018)
21. Haring, K.S., Matsumoto, Y., Watanabe, K.: How do people perceive and trust a lifelike robot. In: *Proceedings of the World Congress on Engineering and Computer Science*, vol. 1 (2013)
22. Haring, K.S., Watanabe, K., Silvera-Tawil, D., Velonaki, M., Takahashi, T.: Changes in perception of a small humanoid robot. In: *2015 6th International Conference on Automation, Robotics and Applications (ICARA)*, pp. 83–89. IEEE (2015)
23. Hinds, P.J., Roberts, T.L., Jones, H.: Whose job is it anyway? A study of human-robot interaction in a collaborative task. *Hum.-Comput. Interact.* **19**(1), 151–181 (2004)
24. Waytz, A., Morewedge, C.K., Epley, N., Monteleone, G., Gao, J.-H., Cacioppo, J.T.: Making sense by making sentient: effectance motivation increases anthropomorphism. *J. Pers. Soc. Psychol.* **99**(3), 410 (2010)
25. Waytz, A., Gray, K., Epley, N., Wegner, D.M.: Causes and consequences of mind perception. *Trends Cogn. Sci.* **14**(8), 383–388 (2010)
26. Constance, S., Shuler, Z., Klaber, L., Rakauskas, M.: Conformity, compliance, and obedience (1999). [Online]. <https://www.units.miamioh.edu/psychersite/cults/ccoshtml>
27. Phillips, E., Zhao, X., Ullman, D., Malle, B.F.: What is human-like?: Decomposing robots’ human-like appearance using the anthropomorphic robot (abot) database. In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 105–113. ACM (2018)
28. Parasuraman, R., Sheridan, T.B., Wickens, C.D.: A model for types and levels of human interaction with automation. *IEEE Trans. Syst. Man Cybern.-Part A: Syst. Hum.* **30**(3), 286–297 (2000)
29. Beer, J.M., Fisk, A.D., Rogers, W.A.: Toward a framework for levels of robot autonomy in human-robot interaction. *J. Hum.-Robot Interact.* **3**(2), 74–99 (2014)
30. Onnasch, L., Wickens, C.D., Li, H., Manzey, D.: Human performance consequences of stages and levels of automation: an integrated meta-analysis. *Hum. Factors* **56**(3), 476–488 (2014)
31. Fessler, D.M., Holbrook, C., Snyder, J.K.: Weapons make the man (larger): formidability is represented as size and strength in humans. *PloS One* **7**(4), e32751 (2012)
32. Satterfield, K., et al.: Investigating compliance in human-robot teaming. In: *2nd International Conference on Intelligent Human Systems Integration (IHSI 2019): Integrating People and Intelligent Systems* (2019)
33. McCambridge, J., Witton, J., Elbourne, D.R.: Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. *J. Clin. Epidemiol.* **67**(3), 267–277 (2014)
34. Wickström, G., Bendix, T.: The Hawthorne effect—what did the original Hawthorne studies actually show? *Scand. J. Work Environ. Health* **26**, 363–367 (2000)

35. Parsons, H.M.: What happened at Hawthorne?: New evidence suggests the Hawthorne effect resulted from operant reinforcement contingencies. *Science* **183**(4128), 922–932 (1974)
36. Lucas, G.M., Gratch, J., King, A., Morency, L.-P.: It’s only a computer: virtual humans increase willingness to disclose. *Comput. Hum. Behav.* **37**, 94–100 (2014)
37. Linder, S.P., Nestruck, B.E., Mulders, S., Lavelle, C.L.: Facilitating active learning with inexpensive mobile robots. *J. Comput. Sci. Coll.* **16**(4), 21–33 (2001)
38. Hashimoto, T., Kato, N., Kobayashi, H.: Development of educational system with the android robot SAYA and evaluation. *Int. J. Adv. Robot. Syst.* **8**(3), 28 (2011)
39. Saerbeck, M., Schut, T., Bartneck, C., Janse, M.D.: Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1613–1622. ACM (2010)
40. Srinivasan, V., Takayama, L.: Help me please: robot politeness strategies for soliciting help from humans. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 4945–4955. ACM (2016)
41. Hayes, C.C., Miller, C.A.: *Human-Computer Etiquette: Cultural Expectations and the Design Implications they Place on Computers and Technology*. CRC Press, Boca Raton (2010)
42. Parasuraman, R., Miller, C.A.: Trust and etiquette in high-criticality automated systems. *Commun. ACM* **47**(4), 51–55 (2004)
43. de Visser, E.J., Pak, R., Shaw, T.H.: From ‘automation’ to ‘autonomy’: the importance of trust repair in human-machine interaction, *Ergonomics* **61**, 1409–1427 (2018)
44. de Visser, E.J., et al.: Almost human: anthropomorphism increases trust resilience in cognitive agents. *J. Exp. Psychol.: Appl.* **22**(3), 331 (2016)
45. Jackson, R.B., Wen, R., Williams, T.: Tact in noncompliance: the need for pragmatically apt responses to unethical commands (2019)