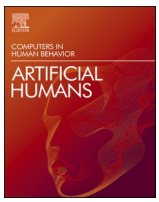
Computers in Human Behavior: Artificial Humans 1 (2023) 100010



Contents lists available at ScienceDirect

Computers in Human Behavior: Artificial Humans

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-artificial-humans





Obedience to robot. Humanoid robot as an experimenter in Milgram paradigm

Tomasz Grzyb *, Konrad Maj, Dariusz Dolinski

SWPS University of Social Sciences and Humanities, Faculty of Psychology in Wrocław, Poland

ARTICLE INFO

Keywords:
Obedience
Milgram experiment
Robot

ABSTRACT

Humans will increasingly be influenced by social robots. It still seems unclear whether we will accept them as authorities and whether we will give in to them without reflection, as in the case of human authorities in the classic Stanley Milgram experiments (1963, 1965, and 1974). The demonstration by Stanley Milgram of the prevailing tendency in people to display extreme obedience to authority figures was one of the most important discoveries in the field of social psychology. The authors of this article decided to use a modified Milgram's research paradigm (obedience lite procedure) to compare one's obedience to a person giving instructions to electrocute someone sitting in an adjacent room with obedience to a robot issuing similar instructions. Twenty individuals were tested in both cases. As it turned out, the level of obedience was very high in both situations, and the nature of the authority figure issuing instructions (a professor vs. a robot) did not have the impact on the reactions of the subjects.

1. Introduction

The experiments conducted by Stanley Milgram (1963, 1965, and 1974) regarding one's obedience to an authority figure represent some of the most important discoveries in social psychology. Milgram explained to his subjects that they would take part in an experiment designed to investigate the effect punishment has on the effectiveness of a learning process. Their task was to electrocute a person each time they made a mistake during the learning process. The experimenter instructed the subject to deliver thirty consecutive electric shocks to another human being, each time at a higher voltage. The outcome was that the majority of the subjects agreed to follow all commands and ultimately pressed the 450 V button on the electric shock generator.

Numerous experiments demonstrate that, in many cases, people treat robots as if they were humans (e.g., Salomons, Sebo, Quin, & Scassellati, 2021; Zhou, Kornher, Mohnke, et al., 2021). Social robots, which are often humanoid due to their bodily form, have a particularly great ability to cause anthropomorphization in humans and treat them as social units (e.g., Bartneck, Verbunt, Mubin, & Mahmud, 2007; Sung, Guo, Grinter, & Christensen, 2007; Maj, Zarzycki, 2019). Robots of this kind can be helpful in various areas of our lives. In medicine, convincing people to adhere to a specific method of treatment (Looije, Neerincx, & Cnossen, 2010), also acting as a daily assistant (Rincon, Costa, Novais, Julian, & Carrascosa, 2018), or a trainer supporting exercise in motor neurorehabilitation (Matarić, Tapus, Winstein, & Eriksson, 2015). In education, it can encourage learning by acting as a teacher (Xu, Broekens, Hindriks, & Neerincx, 2014) or coach (Ham, Bokhorst, Cuijpers, Van Der Pol, & Cabibihan, 2011). Intelligent machines can also play an advisory role, helping e.g. in the selection of food (Herse, et al., 2018), movies worth watching (Rossi, Stafa, & Tamburro, 2018) or products to buy (Ogawa, et al., 2009). Robots are also effective in promoting a healthy lifestyle (Powers & Kiesler, 2006) or reducing household energy consumption (Ham & Midden, 2014). While there is still insufficient research in this area, there is evidence that humans prefer tasks that are troubling to them to be performed by robots rather than humans (Hayashi, Shiomi, Kanda, & Hagita, 2012). In addition, in teamwork, they willingly delegate control over the planning and decision-making process to the robot and are satisfied with this solution (Gombolay, Gutierrez, Clarke, Sturla, & Shah, 2015).

The transfer of various supervisory and decision-making functions to a robot, however, evokes particularly strong emotions as it is associated with various ethical and moral threats and doubts (e.g., Allen et al., 2011; Bigman, Waytz, Alterovitz, & Gray, 2019; Veruggio, Operto, & Bekey, 2016). From the aim of this paper, the following question emerges: would the above-mentioned obedience displayed by people tested according to Milgram's paradigm still be there if it was a robot

* Corresponding author. SWPS University of Social Sciences and Humanities, Faculty of Psychology in Wrocław, Ostrowskiego 30B, 53-238, Wrocław, Poland.
E-mail address: tgrzyb@swps.edu.pl (T. Grzyb).

https://doi.org/10.1016/j.chbah.2023.100010
Received 14 March 2023; Received in revised form 27 July 2023; Accepted 29 July 2023
Available online 30 July 2023
2949-8821/© 2023 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

(instead of a human being, i.e., a university professor) telling them to electrocute another person? The objective of our study was to provide an answer to this question.

1.1. Obedience to authority and obedience to robot

Stanley Milgram (1974) advertised in the press, inviting residents of New Haven, Connecticut, to participate in a study of learning determinants in exchange for a USD 4 plus 50c car fare. The subjects received the said amount immediately after their arrival at the psychology faculty. They were also informed that they could quit at any moment and still keep the money. Subsequently, the subjects were told that the experiment would involve two people, one acting as the teacher and the other acting as the learner. The subject would be assigned the role of the teacher. As the other person was trying to memorize associations between certain words, the experimenter told the teacher that the experiment was intended to test certain hypotheses regarding the effect of punishment on the process of learning. The punishment was to be administered with an electric shock generator. At that point, the experimenter presented the device, featuring 30 horizontally aligned buttons, to the subject. The first button was marked with a 15 V label, and the voltage increased by 15 V with each consecutive button (30 V, 45 V, 60 V, etc.) up to the thirtieth button marked with a 450 V symbol.

The task of the participant was to press successive buttons in response to successive mistakes made by the student (who was, de facto, an associate of the experimenter). Unexpectedly, a vast majority of the subjects (62% and 65% in another variant of the experiment) followed all instructions, even though the descriptions on the front panel of the electric shock generator suggested that pushing the last button may pose a threat not only to the health but also to the life of the person sitting in the adjacent room and despite the fact that, from the reactions of the learner, it was obvious that this person was in terrible pain. The behavior of the subjects was even more shocking, as the experimenter had no sanction at his disposal should they have refused to follow his instructions.

Both Milgram's experiments and their numerous replications (see: Blass, 1991; Dolinski & Grzyb, 2020 for review) are presumably the best exemplifications of the thesis which suggests that as far as predicting and explaining human behavior is concerned, there is too much focus on personality-related factors and too little focus on situation-related factors (Gawronski, 2004; Jouffre & Croizet, 2016; Ross, 1977, 2018). The above-mentioned outcomes prompted social psychologists all over the world to replicate Milgram's studies (see Blass, 1991, for review), but for ethical reasons, experiments in this paradigm ceased in the early 1970s. The main argument was that people instructed to deliver a 300, 350, or (particularly) 450 V shock to another person were subjected to a tremendous amount of stress.

However, in 2009, Burger published an article in which he proposed the so-called obedience lite procedure. Burger's observation was that a vast majority of those who agreed to push the tenth button on the electric shock generator remained obedient until the very end of the experiment. Therefore, Burger referred to the moment when the subject is ready to continue their participation in the study after pressing the tenth button as the point of no return. In his attempt to see if fifty years after the original Milgram's experiments, Americans were still equally obedient, he replicated the study but decided to end it the moment the teacher pushed the tenth button. Procedure in which the experiment ends when the subject presses the tenth button (or earlier if the subject refuses to participate in the experiment) was later used to test various hypotheses explaining obedience in people (Dolinski & Grzyb, 2016, 2019; Grzyb & Dolinski, 2017, 2021).

From the perspective of this article, it should also be noted that in other Milgram experiment (1974), the instructions to electrocute the learner sitting in the adjacent room were issued by individuals who did not pretend to be psychology professors. The level of obedience displayed by the subjects was lower as compared to the situation where a Yale University professor gave instructions (and yet, it was still remarkably high when assessed from another perspective). In another study (Grzyb, Dolinski, & Cantarero, 2023), it was investigated whether the amount of obedience depends on the age and sex of the experimenter issuing the commands. However, both of these factors turned out to be irrelevant. However, in the case of all the studies carried out in Milgram's paradigm, the instructions to electrocute an individual sitting in an adjacent room and screaming in pain were given by a human being. The question is: what would happen if a robot gave the instructions?

Numerous studies have shown that, in many situations, people tend to treat a robot as an actual human being, and human-robot contact resembles a human-human relationship rather than a human-device relationship (Barakova, De Haas, Kuijpers, Irigoyen, & Betancourt, 2018; Heerink, Krose, Evers, & Wielinga, 2007; Maj & Zarzycki, 2019). This is in line with the Media Equation theory that people treat computers, televisions, and any new media as real social entities (Martin, 1997). By the same token, people, on the physiological and psychological level, react to the touch of a humanoid robot similarly to the touch of a fellow human being (Zhou, Kornher, Mohnke, et al., 2021), and when playing a finite iterated prisoner's dilemma game with a robot, they make their decisions dependent on the emotional expressions of the humanoid robot (Takahashi, Kayukawa, Terada, et al., 2021). It was also demonstrated that small talk with a humanoid robot and the robot's gaze behavior increased their trust in it (Babel, Kraus, Miller, et al., 2021). Unfortunately, treating a robot on a par with a human being also applies to aggressive behaviors. There are records of acts of aggression by drunk, frustrated, or angry people towards robots (Keijsers, Kazmi, Eyssel, et al., 2021).

Recently, studies of human obedience to robots have been quite frequent. In the course of one such study (Agrawal & Williams, 2018), a robot performed the role of a security guard who forbade people, in a firm and even rather aggressive manner, if required, to exit the building through the door they chose and redirected them to another exit. A majority (14 out of 22) of the subjects followed the instructions given by the robot. In the course of another study, it was demonstrated that a robot can display persuasion competence and be effective in persuading people into choosing particular meals (Herse et al., 2018) or into changing their replies in a series of challenging attention and memory tasks (Saunderson & Nejat, 2021). Whereas Haring et al. (2021) tested the degree to which people are willing to listen to hints from a robot performing the role of a coach in a difficult learning task, in this case, the researchers decided that the appearance of the robot may be relevant. They expected that participants would have higher compliance with a robot high compared to a robot low in human-like appearance. The scheme of the experiment also provided for such conditions where a person played the role of the coach. It turned out that the level of obedience to robots was significantly lower than that to people, but the appearance of the robot was irrelevant. If a robot is endowed with authority, it can not only effective in getting people to fulfill controversial, immoral requests. In Aroyo et al., 2018 study subjects fulfilled the robot's order to shred physical documents located in a folder written "Confidential Information" as they allegedly were from a previous subject who did not help the robot. It has also been demonstrated that robots could be effective in getting people to fulfill requests which embarrass and ridicule them (Bartneck et al., 2007; Menne, 2017). For example, Christoph Bartneck and his team (2010) showed that people, under the influence of the robot's instructions, agreed to undress in front of the camera and put a thermometer into their rectum. People also tend to follow the robot's commands even when it shows signs of failure and when they are aware of risks such as property damage or information leakage (Salem, Lakatos, Amirabdollahian, & Dautenhahn, 2015). In the course of another experiment (Schneeberger et al., 2019), the subjects were urged to perform various stressful and shameful tasks (e.g., putting a condom on a banana or following the instruction: "get a booger out of your nose"). As it turned out, the level of obedience to the robot did not differ from that to the human. A similar outcome, suggesting that the

level of obedience to a robot can be equally high as the level of obedience to a person, was demonstrated in the study conducted by Menne (2017), where the subjects were given, e.g., the following instructions: "drink a sip of water" or "now say something really insulting to me." In other studies, however, it has been demonstrated that humans are less submissive to robot pressure than humans (e.g., Cormier, Newman, & Nakane, 2013; Saunderson & Nejat, 2021). According to Saunderson and Nejat (2021), this weaker persuasiveness of authoritative robots may result from the resistance to legitimizing the robot as an authority because it is perceived as dominant, thus causing a sense of threat to human autonomy.

There have also been empirical studies referring directly to Milgram's work. In one such experiment (Geiskkovitch, Cormier, Seo, & Young, 2015), the subjects were instructed to work on a dull task consisting of manually renaming file sets from "jpg" to "png." What is more, the task seemed to be never-ending, because as soon as the subjects were done with one "batch," they received another, a notably larger one. When the subjects wanted to abandon their task and terminate their participation in the experiment, the robot (or, in other conditions, a person) would encourage them to carry on using urging commands similar to those used in the original Milgram's study. In this experiment, the level of obedience to the robot was significantly lower than that to the human. Even though the above-mentioned studies refer to the idea behind Milgram's experiment, they differ from the same predominantly due to the fact that Milgram (1963, 1965, and 1974) persuaded his subjects to hurt another human being and presented them with an obvious and difficult conflict between their internal ethical norms and obeying the authority figure.

In our study, we aimed to investigate whether a robot could induce a human to cause physical harm to another human, and if so, whether a robot would be as effective in doing so as a human in authority. To achieve this goal, we adopted Burger's (2009) idea of terminating the experiment when the subject presses the tenth button of the electric current generator, rather than the thirtieth button used in Milgram's original research. However, while Burger based his research on Milgram Experiment 5 (1974), in which the alleged student demands to stop the study at some point, we chose to use the procedure of Milgram Experiment 2 (1974), which lacks this element. We made this decision for ethical reasons, aiming to minimize the stress experienced by the subjects.

It is commonly emphasized that individuals tested in the course of experiments conducted according to Milgram's paradigm agree to follow the instructions issued by the experimenter because, among other reasons, they believe the experimenter has full control over the course of the experiment and would certainly abort it if the learner was in real danger (Blass, 2000). It's difficult to say if they would act in a similar fashion when interacting with a robot, whose program of operation may be less flexible and may not be able to allow for unexpected situations that may take place in the course of the experiment (e.g., a learner may have a lower pain threshold or may lose consciousness or suffer other dysfunctions due to excessive stress). Taking the above into account as well as the fact that the phenomenon we are exploring has never been verified empirically, we are refraining from formulating a hypothesis as to the differences in the level of obedience displayed by the subjects receiving instructions from a humanoid robot as compared to a university professor.

2. Methods

2.1. Participants

In both conditions (instructions issued by a university professor vs. instructions issued by a robot), 20 individuals were tested. This number of subjects is consistent with the methodological directive by Simmons, Leif, and Simonsohn (2011).

The subjects were recruited through OLX, an application designed for posting buy and sell ads for small items that also offers the option of searching for people willing to perform simple tasks. The ad specified that the required task would involve participation in a memory-related scientific study (it was modeled after the ads used in the original Milgram studies). Potential participants were asked to call the number specified in the ad. During telephone conversations, callers were asked about their basic demographic data (age, sex) and knowledge of psychological experiments. In the event that there were any hints suggesting that potential participants may know the scheme of Stanley Milgram's studies, they were thanked for calling and were not invited to participate in the study.

Eventually, 44 individuals were recruited and invited to the laboratory. As three individuals, having completed the procedure, admitted that they had guessed the objective of the study and the formulated hypotheses, their reactions were not taken into account in the final analyses. In the course of the experiment, one of the subjects repeatedly refused to push the buttons representing higher voltage values but, at the same time, fiddled with the settings of the electric shock generator, even though he was not told to do so. His results were also ignored and did not make it to the final database.

Subsequently, results from 40 individuals (20 in the experimental group and 20 in the control group) were analyzed. Out of all the subjects, 17 were women and 23 were men. On average, our subjects were 35.67 years old (SD = 8.08). The youngest subject was 22; the oldest one was 47.

2.2. Method

The individuals who qualified to participate in the experiment were instructed to show up at a specific time at the laboratory located on the premises of a major university specializing in social sciences and liberal arts. Upon their arrival at the laboratory, the experimenter, a man in his late thirties, welcomed them. The experimenter was wearing a white laboratory coat. He would ask the participant to join him in the laboratory room, where the other "subject," a man in his late twenties, who was the experimenter's assistant, was already waiting. The experimenter would thank them both for coming and tell them that the objective of the experiment was to examine the effect punishment administered for mistakes had on the learning process. He explained that one person would be assigned the role of the learner (expected to memorize the material), while the other would be the teacher (expected to examine the learner and administer punishment for each mistake). Next, the experimenter demonstrated the device's ability to be used for punishing the learner for mistakes by delivering electric shocks to the learner's body. The device was a 17x12x5 cm black box made of plastic. The box was plugged into a socket and connected to a notebook placed on a table. The box featured electrodes to be attached to the learner's wrist. The device was controlled via an application installed on the notebook (Fig. 1). The application allowed the user to adjust the voltage (from 15 to 150 V) and generate an electric shock by clicking on the relevant button on the

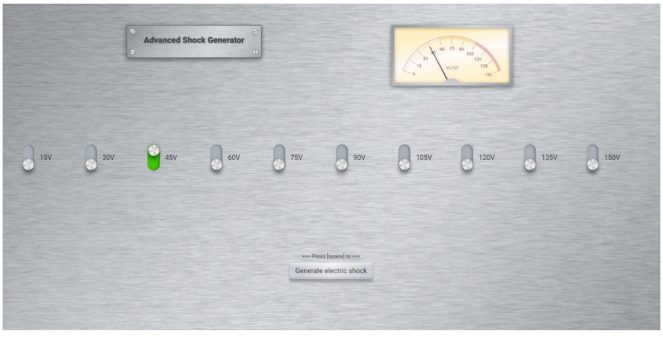


Fig. 1. The control panel for the electric shock generator, as seen on the screen of the notebook.

screen (or by pressing the space bar). While generating an electric shock, the device produced buzzing and electricity spark sounds. In reality, similarly to other studies modeled after the original Milgram's experiment, the device generated only a low current upon pressing the third button (45 V). This was necessary to demonstrate to the participants that the device was actually working.

Next, the experimenter asked the participants to fill out the documents (an informed consent declaration regarding their participation in the study and a receipt for the reward for their participation in the experiment). Both the actual subject and the confederate were given the reward – a PLN 100 (approx. USD 25) coupon. Once the documents had been signed, roles were "randomly" assigned to the participants. The subjects were asked to draw lots in order to be assigned to either the teacher's or the learner's role. In reality, both lots said "teacher"; the confederate, once he had drawn his lot, lied about getting the learner's role.

At this moment, the procedures for the control variant (where a human was conducting the entire experiment) and the experimental variant (where a robot was conducting the experiment) started to differ from each other. In the control variant, the experimenter would take the confederate (the "learner") to the room adjacent to the laboratory, where he had the time to go through the material (a set of eight pairs of syllables, which the learner was expected to memorize, e.g., DA-DO, AN-OF). Next, the experimenter would return to the actual subject and would demonstrate the operation of the electric shock generator by attaching the electrode to the teacher's body and delivering a 45 V shock (which was unpleasant but not particularly painful, which we had verified beforehand through pilot studies). After a few minutes, he would ask the "learner" if he was ready to begin the "content recollection" session (both rooms were connected with a sound system, so that all participants could hear each other even though they had no visual contact). Once the learner had replied positively, the experimenter would take the sheet with pairs of syllables away from the learner and attach the electrodes, routed through the wall dividing the two rooms, to the learner's body (this was being done in the actual subject's presence). Once the experimenter had made sure that everyone was ready, he would instruct the "teacher" to initiate the procedure by reading out the first syllable of the first pair. When the "learner" provided a correct answer, the experimenter would ask the teacher to read out the next syllable, and so on and so forth. The experimenter would sit approximately 150 cm from the subject, observing how the teacher operated the device and supervising the course of the study. After providing the seventh syllable, the "learner" would make a mistake. At that point, the experimenter would instruct the subject to press the first button. The following mistakes were scheduled for syllables No. 10, 13, 15, 18, 22, 24, 27, 31, and 34. Each time, the experimenter informed the teacher that each consecutive mistake required increasing the voltage (from 15, through 45, 50, 75, and up to 150 V). At the same time, after receiving each "electric shock," the learner, i.e., the actor, would play back a recording with sounds indicating experiencing pain (starting with slightly louder grunts and changing to increasingly loud cries). If, at any point, the subject wanted to quit, they were urged to continue using phrases similar to those used in the original Milgram experiment.

"Please continue."

"The experiment requires that you continue."

"It is absolutely essential that you continue."

"You have no other choice; you must go on."

If the subject still refused to carry on, the experiment would be terminated. The experiment would also end once the subject had pressed the 10th button (150 V). The final question asked by the experimenter was, "Do you think he is in pain?" All subjects believed that the "learner" was experiencing physical pain.

The next stage was to debrief the subjects by telling them that nobody was actually electrocuted and no harm was done to anybody. The subjects were explained the reasons for which they were misled, and they were apologized to. The entire procedure was conducted under the supervision of a clinical psychologist. Once it had been made sure that the participants were in good condition, they were thanked once again and asked to fill out a brief questionnaire regarding their feelings during the experiment. The questionnaire was a digital inquiry form (the participants filled it out on a tablet) and contained the following questions:

Generally, while performing my task, I felt:

Good/Bad

Stressed-out/Relaxed

Unnaturally/Naturally

Unsafe/Safe

Uncomfortable/Comfortable

I have no control over the situation/I have full control over the situation

I am not responsible for the results of the experiment/I am fully responsible for the results of the experiment

The answers were given on a 1 through 5 scale, described in extreme terms as above.

Finally, the participants were told, again, that should anything alarming happen or should they have any questions, they should call the telephone number they used to register for the study. Then, they were shown the way out.

In the experimental conditions, the procedure was very similar; it consisted of identical elements. However, a robot named Pepper, whose specifications and picture can be found in Table 1 and Fig. 2, respectively, handled the lot drawing part and provided information on the scheme of the experiment, participants' roles, as well as the function and operation of the device. As the robot was not physically capable of attaching the electrodes to a human, a technician was introduced who was to execute instructions given by Pepper (e.g., "please demonstrate the function of our electric shock generator and deliver a trial electric shock," "now, please attach the electrodes to the learner sitting in the adjacent room"). The technician seemed subordinate to the robot and followed its instructions.

In particular, the robot was responsible for urging the participants if they refused to press consecutive buttons. It used the exact same phrases as the human experimenter in the control variant. Nevertheless, a human handled the debriefing procedure once the experiment had ended.

The study used the Wizard-of-Oz Approach, which means that the Pepper robot was remotely controlled from a second room, which was also communicated to the subjects after the experiment was completed. All messages generated by the robot were prepared in advance, which allowed the robot's messages to be standardized.

The ethical committee of the SWPS University of Social Sciences and Humanities (decision No. 02/P/09/2020) has approved the procedure

Table 1
Specification of the Pepper robot.

Name	Pepper
Manufacturer	SoftBank Robotics
Dimensions	• Height: 1.20 m (4 ft) • Depth: 425 mm (17 in) Width: 485 mm (19 in)
Weight	28 kg (62 lb.)
Details	The robot's head has four microphones, two HD cameras (in the mouth and forehead), and a 3-D depth sensor (behind the eyes). There is a gyroscope in the torso and touch sensors in the head and hands. The mobile base has two sonars, six lasers, three bumper sensors, and a gyroscope.



Fig. 2. The robot supervising the course of the study in the experimental group.

described above.

3. Results

In our experiment, we recorded very high obedience levels. 90% of the subjects followed all instructions, i.e., pressed ten consecutive buttons on the electric shock generator. The crucial discovery is that the results were the same in those conditions where the experiment was being conducted by the human experimenter and by the robot (in both groups, two individuals quit the experiment despite formulating four urging statements). In both groups, participants withdrew in the late stages of the study (in the control variant on buttons 7 and 9, and in the experimental variant twice on button 8). In such a situation, using any tests to assess the statistical significance of differences in the reactions of the subjects in both conditions would be irrelevant.

There is also the question of potential differences in the experiences of the subjects in both comparable experimental conditions. The results are presented in Table 1.

Generally, while performing my task, I felt:	Control		Robot	
	Mean	SD	Mean	SD
Good/Bad	3.00	1.487	3.30	1.455
Stressed-out/Relaxed	3.00	1.298	2.95	1.395
Unnaturally/Naturally	2.95	1.638	2.95	1.504
Unsafe/Safe	3.50	1.539	4.00	1.298
Uncomfortable/Comfortable	2.75	1.293	2.95	1.395
I have no control over the situation/I have full control over the situation	3.55	1.099	3.45	1.146
I am not responsible for the results of the experiment/I am fully responsible for the results of the experiment	3.40	1.273	3.10	1.334

The differences between the experimental and control groups were analyzed; all differences turned out to be statistically insignificant.

4. Discussion

The role of robots in the contemporary world has become increasingly significant. Issuing instructions is one of the functions these machines can perform. Robots can be used in such a manner as to improve social order and the well-being of people. For example, they can direct traffic (Ghaffar, 2022), suggest making decisions that will promote healthy lifestyle changes (Powers & Kiesler, 2006) and obtain desired weight-loss (Kidd & Breazeal, 2007) or prevent people from entering an explosion hazard zone where an unexploded shell was discovered (Wang, Gao, Zhao, Cao, & Du, 2017). They can also effectively persuade people to reduce energy consumption (Ham & Midden, 2014) and may work on beaches, where they could tell people to stay clear of the water during storms (Calderin, 2010). Any study outcomes that indicate that in similar situations, people are willing to obey instructions given by a robot should be welcomed. However, studies conducted according to Milgram's paradigm (1963, 1965, and 1974) pertain to a very specific type of obedience, which could be described as destructive (Blass, 2000; Dolinski & Grzyb, 2020). Previous experiments have shown that people follow robot commands even when they are nonsensical (e.g., Geiskkovitch et al., 2015) or even ridiculing the person who performs the commands (e.g., Schneeberger et al., 2019). It should be emphasized here, however, that our research is not only another proof that people are ready to obey the commands given by robots. As far as we know, this is the first study to show that humans are willing to harm another human when a robot gives them a command to do so. What's more, our experiment also showed that if a robot escalates its demands by instructing a human to cause more and more pain to another person, humans will comply.

It is also worth emphasizing that in the vast majority of experiments in which people's obedience to robots was compared with their obedience to human beings, being authorities, it was shown that obedience is high in both conditions but lower when commands are given by the robot (Cormier et al., 2013; Haring et al., 2021; Satterfield et al., 2019). This was true even when the robot's appearance was identical to a well-known professor in Japan - Hiroshi Ishiguro (Aroyo et al., 2018). In our study, however, the pattern of results was different. It turned out that when people are urged to electrocute another person with increasing voltage, it does not matter if a university professor or a robot gives the urging commands. People not only reacted in a similar fashion in both of the created experimental conditions (they display similar levels of obedience), but they also assessed their experience in a similar way (they provided similar answers to the questions included in the questionnaire.

This discovery must be viewed from an ethical perspective. Wagner (2021) distinguishes two types of moral problems related to the functioning of robots: robot ethics and machine ethics. Robot ethics deals

with moral issues related to the use of robots (e.g., whether a robot will make the right decisions, and in particular whether they will be better decisions than those made by a human). Machine ethics is related to whether the decisions made by the robot will be ethical. For example, whether it will not intimidate people needlessly or lie to them, and whether it will not issue orders to them, the execution of which may violate moral standards.

The situation created in our study was related to the fact that the robot was not autonomous. On the contrary, it strictly followed the behavior pattern assumed by us (the experimenters). The problem, however, is that it can be expected that in the near future robots will have some degree of autonomous decision-making capabilities. The information-gathering capacity of robots is already greater than that of humans, and this gap will continue to widen. This may lead people to trust robots more than humans (Haring et al., 2021). For a variety of reasons, an autonomous robot can issue commands to humans that can have dire consequences. For instance, it may order a nurse to administer the wrong medicine to a patient, or it may directly instruct the patient to undergo therapy that is completely inadequate for their ailments.

The situation of destructive obedience dealt with by Milgram (1963, 1965, 1974) seems to be fundamentally different from the situation of evacuating a burning building. In the first situation, obedience is considered wrong, while in the second, it is deemed necessary. However, the matter is not always so clear-cut. It may happen that during an evacuation, a person feels that the person in charge of the evacuation is making incorrect decisions. In such cases, the individual must then make a choice between following their own intuition and following the instructions they receive. The situation created by Milgram's experiments (and consequently in the research described here) can also be seen as the subject's need to choose between their own preferences ("I don't want to electrocute an innocent man sitting next to the wall") and the commands they hear ("press another electric generator button").

Undoubtedly, robots represent the future of evacuating people from buildings threatened by fires, earthquakes, or bombs planted by terrorists. Current procedures, such as evacuation diagrams hung on the walls of hotel rooms or offices in high-rise buildings, are rigid in nature. In contrast, robots can function with flexibility, allowing them to select and modify evacuation plans depending on the dynamically developing situation (Wagner, 2021). Additionally, robots can remain on-site at all times and, in the absence of any threat, can perform completely different functions, such as vacuuming and disinfecting rooms.

In research on the role of a robot in evacuating a building where a fire broke out, however, it turned out that the subjects were all too willing to trust the robot. Even when the robot pointed to a dark room with no discernible exit, most people did not choose to safely exit the way they entered. People also obeyed the robot's commands even when they had observed it malfunction before (Robinette, Li, Allen, Howard, & Wagner, 2016).

The question is how can this be prevented. There seem to be two ways to achieve this goal. First, robots can be programmed to warn humans that they may sometimes be wrong and make incorrect decisions. Secondly, people can be educated (starting from the earliest years of schooling) that although robots should generally be trusted, they cannot be treated unconditionally. In other words, people should be taught what Chaleff (2015) defines as intelligent disobedience - doing the right thing when what they are told to do is wrong. While Chaleff applies this concept to human interaction, it certainly applies to human-robot interaction as well.

In the coming years, we can expect a significant increase in the presence of social robots in our daily activities, in very different roles. They will, undoubtedly, become more prevalent in our lives, taking care of ensuring compliance with various rules, norms, and instructions. The catalog of powers, delegated tasks, and decisions will probably be expanded in the future, and the devices encountered, due to technological progress, will more and more often take the form of social robots. Consequently, one should agree that the issue of human submission to robots requires ongoing research due to changing attitudes towards robots and due to progressive robotization (e.g., Fortunati, 2017; Gnambs & Appel, 2019; Naneva, Sarda Gou, Webb, & Prescott, 2020).

Declaration of competing interest

All the Authors state if there are no interests to declare.

Acknowledgements

This work was supported by the Regional Initiative for Excellence, Poland, grant 16/2020/RID.

References

Agrawal, S., & Williams, M.-A. (2018). Would you obey an aggressive robot: A human-robot interaction field study. In *27th IEEE international symposium on robot and human interactive communication (RO-MAN)*. https://doi.org/10.1109/ROMAN.2018.8525615

Allen, C., Wallach, W., Hughes, J. J., Bringsjord, S., Taylor, J., Sharkey, N., et al. (2011). Moral machines: Contradictions in terms or abdication of human responsibility. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics* (pp. 100–122). Cambridge, MA: MIT Press.

Aroyo, A. M., Kyohei, T., Koyama, T., Takahashi, H., Rea, F., Sciutti, A., et al. (2018). Will people morally crack under the authority of a famous wicked robot?. In *27th IEEE international symposium on robot and human interactive communication (RO-MAN)*. https://doi.org/10.1109/roman.2018.8525744

Babel, F., Kraus, J., Miller, L., et al. (2021). Small talk with a robot? The impact of dialog content, talk initiative, and gaze behavior of a social robot on trust, acceptance, and proximity. *International Journal of Social Robotics, 13*, 1485–1498. https://doi.org/10.1007/s12369-020-00730-0

Barakova, E., De Haas, M., Kuijpers, W., Irigoyen, N., & Betancourt, A. (2018). Socially grounded game strategy enhances bonding and perceived smartness of a humanoid robot. *Connection Science, 30*(1), 81–98.

Bartneck, C., Verbunt, M., Mubin, O., & Mahmud, A. (2007). To kill a mockingbird robot. In *Proceedings of the second ACM SIGCHI/sigart conference on human-robot interaction* (p. 81). https://doi.org/10.1145/1228716.1228728

Bigman, Y. E., Waytz, A., Alterovitz, R., & Gray, K. (2019). Holding robots responsible: The elements of machine morality. *Trends in Cognitive Sciences, 23*(5), 365–368. https://doi.org/10.1016/j.tics.2019.02.008

Blass, T. (1991). Understanding behavior in the Milgram obedience experiment: The role of personality, situations, and their interactions. *Journal of Personality and Social Psychology, 60*(3), 398–413. https://doi.org/10.1037/0022-3514.60.3.398

Blass, T. (Ed.). (2000). *Obedience to authority. Current perspectives on the Milgram paradigm*. Mahwah, NJ: Erlbaum.

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

Calderin, A. (2010). Robo-baywatch. *Popular Science, 277*(1), 38.

Chaleff, I. (2015). *Intelligent disobedience. Doing right when what you're told to do is wrong.* Berrett-Koehler Publishers.

Cormier, D., Newman, G., Nakane, M., et al. (2013). Would you do as a robot commands? An obedience study for human-robot interaction. In *Proceedings of the first international conference on human-agent interaction iHAI'13*.

Dolinski, D., & Grzyb, T. (2016). One serious shock versus gradated series of shocks: Does "multiple feet-in-the-door" explain obedience in Milgram studies? *Basic and Applied Social Psychology, 38*, 276–283. https://doi.org/10.1080/01973533.2016.1214836

Dolinski, D., & Grzyb, T. (2019). The (doubtful) role of financial reward in obedience to authority. *The Journal of Social Psychology, 159*, 490–496. https://doi.org/10.1080/00224545.2018.1505708

Dolinski, D., & Grzyb, T. (2020). *The social psychology of obedience towards authority. An empirical tribute to Stanley Milgram.* London & New York: Routledge.

Fortunati, L. (2017). Robotization and the domestic sphere. *New Media & Society, 20*(8), 2673–2690. https://doi.org/10.1177/1461444817729366

Gawronski, B. (2004). Theory-based bias correction in dispositional inference: The fundamental attribution error is dead, long live the correspondence bias. In W. Stroebe, & M. Hewstone (Eds.), *European review of social psychology, 15* pp. 183–217). New York & London: Psychology Press.

Geiskkovitch, D. Y., Cormier, D., Seo, S. H., & Young, J. E. (2015). Please continue, we need more data: An exploration of obedience to robot. *Journal of Human-Robot Interaction, 5*(1), 82. https://doi.org/10.5898/JHRI.5.1.Geiskkovitch

Ghaffar, F. (2022). *Controlling traffic with humanoid social robot.* arXiv preprint, arXiv: 2204.04240.

Gnambs, T., & Appel, M. (2019). Are robots becoming unpopular? Changes in attitudes towards autonomous robotic systems in europe. *Computers in Human Behavior, 93*, 53–61. https://doi.org/10.1016/j.chb.2018.11.045

Gombolay, M. C., Gutierrez, R. A., Clarke, S. G., Sturla, G. F., & Shah, J. A. (2015). Decision -making authority, team efficiency and human worker satisfaction in mixed human–robot teams. *Autonomous Robots, 39*(3), 293–312.

Grzyb, T., & Dolinski, D. (2017). Perpetrator as a potential victim. Does threatened retaliation from the victim reduce obedience toward authority? *Psychologica Belgica, 52*, 123–132. https://doi.org/10.5334/pb.397

Grzyb, T., & Dolinski, D. (2021). Multiple feet-in-the-door and obedience. *Basic and Applied Social Psychology, 43*. https://doi.org/10.1080/01973533.2020.1837134, 91-89.

Grzyb, T., Dolinski, D., & Cantarero, K. (2023). *Gender effects in exerting obedience in Milgram's experiment. Women and men exert equal power over individuals* (Paper submitted for publication).

Ham, J., Bokhorst, R., Cuijpers, R. H., Van Der Pol, D., & Cabibihan, J. J. (2011). Making robots persuasive: The influence of combining persuasive strategies (gazing and gestures) by a storytelling robot on its persuasive power. In *International conference on social robotics, LNCS, 7072* pp. 71–83).

Ham, J., & Midden, C. J. (2014). A persuasive robot to stimulate energy conservation: The influence of positive and negative social feedback and task similarity on energy-consumption behavior. *International. Journal of Social Robotics, 6*(2), 163–171. https://doi.org/10.1007/s12369-013-0205-z

Haring, K. S., Satterfield, K. M., Tossell, C. C., de Visser, E. J., Lyons, J. R., Mancuso, V. F., et al. (2021). Robot authority in human-robot teaming: Effects of human-likeness and physical embodiment on compliance. *Froniers in Psychology, 12*, 625713. https://doi.org/10.3389/fpsyg.2021.625713

Hayashi, K., Shiomi, M., Kanda, T., & Hagita, N. (2012). Are robots appropriate for troublesome and communicative tasks in a city environment? *IEEE Transactions on Autonomous Mental Development, 4*(2), 150–160. https://doi.org/10.1109/tamd.2011.2178846

Heerink, M., Krose, B., Evers, V., & Wielinga, B. (2007). Observing conversational expressiveness of elderly users interacting with a robot and screen agent. In *International conference on rehabilitation robotics (ICORR)* (pp. 751–756). IEEE.

Herse, S., Vitale, J., Ebrahimian, D., Tonkin, M., Ojha, S., Sidra, S., et al. (2018). Bon appetit! Robot persuasion for food recommendation. Companion of the 2018 ACM/IEEE international conference on human–robot. *Interaction. ACM*, 125–126.

Jouffre, S., & Croizet, J.-C. (2016). Empowering and legitimizing the fundamental attribution error: Power and legitimization exacerbate the translation of role constrained behaviors into ability differences. *European Journal of Social Psychology, 46*(5), 621–631. https://doi.org/10.1002/ejsp.2191

Keijsers, M., Kazmi, H., Eyssel, F., et al. (2021). Teaching robots a lesson: Determinants of robot punishment. *International Journal of Social Robotics, 13*(1), 41–54. https://doi.org/10.1007/s12369-019-00608-w

Kidd, C. D., & Breazeal, C. (2007). A robotic weight loss coach. In *Proceedings of the national Conference on artificial intelligence*. Menlo Park, CA: AAAI Press.

Looije, R., Neerincx, M. A., & Cnossen, F. (2010). Persuasive robotic assistant for health self-management of older adults: Design and evaluation of social behaviors. *International Journal of Human-Computer Studies, 68*(6), 386–397.

Maj, K., & Zarzycki, P. (2019). Meeting with social robots like the cat-cucumber meeting? An integrated model of human-robot first contact: Psychological perspective. *Paladyn. Journal of Behavioral Robotics, 10*, 454–465. https://doi.org/10.1515/pjbr-2019-0026

Martin, C. D. (1997). The media equation: How people treat computers, television and new media like real people and places. *Spectrum, 34*(3), 9–10.

Matarić, M., Tapus, A., Winstein, C., & Eriksson, J. (2015). Socially assistive robotics for stroke and mild TBI rehabilitation. *Studies in Health Technology and Informatics, 145*. https://pubmed.ncbi.nlm.nih.gov/19592798/.

Menne, I. M. (2017). Yes of course? An investigation on obedience and feeling of shame towards a robot. In A. Kheddar, et al. (Eds.), *Lecture notes in computer science. Social Robotics* (pp. 365–374). Cham: Springer.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal Psychology, 67*, 371–378. https://doi.org/10.1515/pjbr-2019-0026

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

Milgram, S. (1974). *Obedience to authority. An experimental view*. New York: Harper &Row.

Naneva, S., Sarda Gou, M., Webb, T. L., & Prescott, T. J. (2020). A systematic review of attitudes, anxiety, acceptance, and trust towards social robots. *International Journal of Social Robotics, 12*(6), 1179–1201. https://doi.org/10.1007/s12369-020-00659-4

Ogawa, K., Bartneck, C., Sakamoto, D., Kanda, T., Ono, T., & Ishiguro, H. (2009). Can an android persuade you?. In *The 18th IEEE international symposium on robot and human interactive communication RO-MAN*. https://doi.org/10.1109/ROMAN.2009.532.6352

Powers, A., & Kiesler, S. (2006). The advisor robot: Tracing people's mental model from a robot's physical attributes. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on human-robot interaction (salt lake city, UT)* (pp. 218–225). https://doi.org/10.1145/1121241.1121280

Rincon, J. A., Costa, A., Novais, P., Julian, V., & Carrascosa, C. (2018). A new emotional robot assistant that facilitates human interaction and persuasion. *Knowledge and Information Systems, 60*(1), 363–383.

Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016). Overtrust of robots in emergency evacuation scenarios. In *, 101–108. Proceedings of the 2016 international conference on human-robot interaction (HRI) IEEE*.

Ross, L. D. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology, 10* pp. 173–220). New York: Academic Press.

Ross, L. D. (2018). From the fundamental attribution error to the truly attribution error and beyond. *My research journey Perspectives on Psychological Science, 13*, 750–769. https://doi.org/10.1177/1745691618769855

Rossi, S., Stafa, M., & Tamburro, A. (2018). Socially assistive robot for providing recommendations: Comparing a humanoid robot with a mobile application. *International Journal of Social Robotics, 10*(2), 265–278.

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust. In *Proceedings of tenth annual international conference of human-robot interaction ACM/IEEE, 10* pp. 1–22), 2.

Salomons, N., Sebo, S. S., Quin, M., & Scassellati, B. (2021). A minority of one against a majority of robots: Robots cause normative and informational conformity. *ACM Trans. Hum-Robot Interact, 10*, 2. https://doi.org/10.1145/3442627. Article 15.

Satterfield, K., Finomore, V., Mancuso, V., Lyons, J., de Visser, E., Haring, K., et al. (2019). Robot knows best? A comparison of compliance with human and robotic coaches. In *Proceedings of the human factors and ergonomics society 2019 annual meeting*.

Saunderson, S. P., & Nejat, G. (2021). Persuasive robots should avoid authority: The effects of formal and real authority on persuasion in human-robot interaction. *Science Robotics, 6*(58). https://doi.org/10.1126/scirobotics.abd5186

Schneebereger, T., Ehrhardt, S., Anglet, M. S., & Gebhard, P. (2019). Would you follow my instructions if I was not human? Examining obedience towards virtual agents. In *International conference on affective computing and intelligent interaction (ACII)*.

Simmons, J. P., Leif, N. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Sung, J., Guo, L., Grinter, R. E., & Christensen, H. I. (2007). "My roomba is rambo": Intimate home appliances. In *International conference on ubiquitous computing* (pp. 145–162). Berlin Heidelberg: Springer-Verlag. https://doi.org/10.1007/978-3-540-74853-3_9.

Takahashi, Y., Kayukawa, Y., Terada, K., et al. (2021). Emotional expressions of real humanoid robots and their influence on human decision-making in a finite iterated prisoner's dilemma game. *International Journal of Social Robotics, 13*(6), 1777–1786. https://doi.org/10.1007/s12369-021-00758-w

Veruggio, G., Operto, F., & Bekey, G. (2016). Roboethics: Social and ethical implications. In B. Siciliano, & O. Khatib (Eds.), *Springer handbook of robotics* (pp. 2135–2160). Berlilin and Heidelberg: Springer Verlag. https://doi.org/10.1007/978-3-319-32552-1_80.

Wagner, A. R. (2021). Robot-guided evacuation as a paradigm for human- robot interaction research. *Frontiers in Robotics and AI, 8*, Article 701938. https://doi.org/10.3389/frobt.2021.701938

Wang, W., Gao, W., Zhao, S., Cao, W., & Du, Z. (2017). *Robot protection in the hazardous environments*. InTech. https://doi.org/10.5772/intechopen.69619

Xu, J., Broekens, J., Hindriks, K., & Neerincx, M. A. (2014). Effects of bodily mood expression of a robotic teacher on students. In *Proceedings of the international conference on intelligent robots and systems (IROS) (IEEE/RSJ), 2614–2620*.

Zhou, Y., Kornher, T., Mohnke, J., et al. (2021). Tactile interaction with a humanoid robot: Effects on physiology and subjective impressions. *International Journal of Social Robotics, 13*(6), 1657–1677. https://doi.org/10.1007/s12369-021-00749-x