Piggybacking Robots

Human-Robot Overtrust in University Dormitory Security

Serena Booth
Harvard Paulson SEAS

James Tompkin
Brown University

Hanspeter Pfister
Harvard Paulson SEAS

Jim Waldo
Harvard Paulson SEAS

Krzysztof Gajos
Harvard Paulson SEAS

Radhika Nagpal
Harvard Paulson SEAS

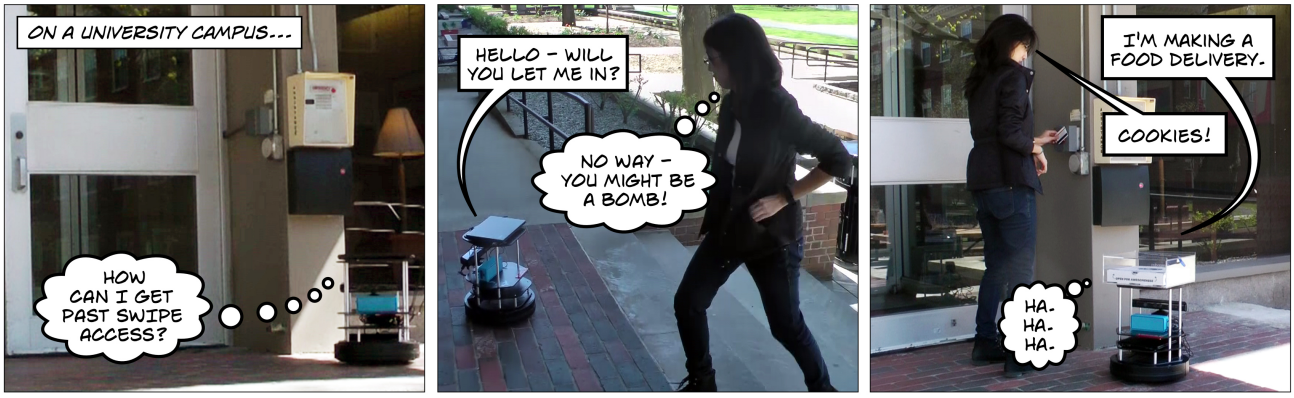

Figure 1: A TurtleBot tries to gain access to a secure facility with an ingenious plan.

ABSTRACT

Can *overtrust* in robots compromise physical security? We positioned a robot outside a secure-access student dormitory and made it ask passersby for access. Individual participants were as likely to assist the robot in exiting the dormitory (40% assistance rate, 4/10 individuals) as in entering (19%, 3/16 individuals). Groups of people were more likely than individuals to assist the robot in entering (71%, 10/14 groups). When the robot was disguised as a food delivery agent for the fictional start-up *Robot Grub*, individuals were more likely to assist the robot in entering (76%, 16/21 individuals). Lastly, participants who identified the robot as a bomb threat demonstrated a trend toward assisting the robot (87%, 7/8 individuals, 6/7 groups). Thus, overtrust—the unfounded belief that the robot does not intend to deceive or carry risk—can represent a significant threat to physical security at a university dormitory.

ACM Reference format:
Serena Booth, James Tompkin, Hanspeter Pfister, Jim Waldo, Krzysztof Gajos, and Radhika Nagpal. 2017. Piggybacking Robots. In *Proceedings of Conference on Human-Robot Interaction, Vienna, Austria, March 6–9, 2017 (HRI '17),* 9 pages.
DOI: http://dx.doi.org/10.1145/2909824.3020211

HRI '17, Vienna, Austria

DOI: http://dx.doi.org/10.1145/2909824.3020211

1 INTRODUCTION

Many offices and residences have physical security systems which restrict access to only authorized individuals. These buildings are threatened by *piggybacking* and *tailgating*, where authorized individuals are followed by unauthorized or unidentified individuals into a secure area, either with or without the consent of the authorized individuals. Piggybacking is common [17], is regularly discouraged [2, 4], and may be actively prevented with patented security systems [19, 24].

With robots assuming roles of increasing autonomy and importance in our daily lives, this problem will expand to include piggybacking robots. A rogue robot can threaten privacy and secrecy, and cyber and physical systems. Humans may *overtrust* an unidentified robot. This effect has been previously demonstrated in the domain of emergency response [20]; however, to our knowledge, human overtrust in granting a robot physical access to a secure building is untested in the literature.

We explored the question of piggybacking robots at university student dormitories (Fig. 1). We placed a robot at a secure-access doorway and had it ask passersby—either individuals or groups—to assist it to enter. To assess the effect of the secure access restriction, we compared the robot asking to enter against asking to exit the building. Then, to increase the apparent legitimacy of the robot, we disguised it as an agent of a fictional food delivery start-up *Robot Grub*, and compared the likelihood of assistance when entering the secure-access building to that of the unmodified robot. Further, we asked participants about their perception of the robot's autonomy to evaluate the relationship between trust and autonomy.

We discovered that individual participants were as likely to assist the unmodified robot to exit (40%, 4/10 individuals) as to enter (19%, 3/16 individuals); this difference was not statistically significant. Groups of people were statistically more likely than individuals to assist the unmodified robot in entering (71%, 10/14 groups). Individual participants were significantly more likely to assist the robot to enter when it had apparent Robot Grub legitimacy (76%, 16/21 individuals) compared to when it did not (19%, 3/16 individuals). Participant perception of robot autonomy was not a predictor of assistance. Lastly, participants who identified the robot as a bomb threat demonstrated a trend toward assisting the robot (87%, 7/8 individuals, 6/7 groups). With these results, we provide evidence that overtrust in robots can represent a significant threat to the physical security of a university dormitory.

2 BACKGROUND

2.1 Defining and Measuring Trust in HRI

The question of how to best define and measure trust is unresolved within the human-robot interaction (HRI), human-computer interaction (HCI), and automation communities. Lee and See define trust in automation to be "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [13]. However, this is more relevant to human-assisting machines than to our scenario, as our focus is not on an individual's goals, but on the ability of the robot to penetrate security. More suitable is Wagner and Arkin's definition: "a belief, held by the trustor, that the trustee will act in a manner that mitigates the trustor's risk in a situation in which the trustee has put its outcomes at risk" [26]. From this, we define *overtrust* to be, "a belief, held by the trustor, that the trustee will not act with deception, and that the trustee will not put the trustor at risk."

Within HRI, we find a division. Desai et al. suggest that trust should be self-reported by participants in human-robot interactions [9], while Salem et al. define trust in terms of compliance with robot instructions [23]. To accommodate the differing ideologies, we measure trust using both techniques, in line with other recent studies of HRI trust [20, 23]. We measure whether or not participants comply with robot instructions, and we also ask participants to rate their trust in autonomous machines in general using a 5-point Likert Scale. While a Likert Scale may not encompass all interpretations of 'trust,' this scale allows participants flexibility in evaluating a complex topic. Further, other researchers have begun to assess trust longitudinally [9, 28].

2.2 Overtrusting Machines

Robinette et al. showed that humans can place too much trust in an emergency response robot: in their study, a simulated emergency occurs and the robot ostensibly leads study participants to safety. Of 26 participants, all followed the emergency robot in this time-critical scenario, even participants who had previously been guided by the robot on obviously indirect, circuitous paths in a non-emergency scenario [20]. When questioned about why they followed the robot's directives despite their own awareness of a straightforward exit route, several participants cited the robot's outward appearance, as a large sign on the robot lit up during the evacuation: "EMERGENCY GUIDE ROBOT." Participants claimed this sign increased the robot's apparent legitimacy.

Apparent legitimacy and trust was further shown to be related by Salem et al. in home companionship scenarios [23]. In their experiment, participants were asked by a robot to perform unusual tasks—such as throwing away private letters or pouring orange juice into a flower pot—which decreased the robot's apparent legitimacy. Participants were statistically less likely to perform irrevocable tasks, and no participants engaged in tasks which breached digital privacy. Further, people's likeliness to trust a robot may be related to its apparent anthropomorphism, as Waytz et al. showed for autonomous vehicles: in their study, participants trusted that an autonomous vehicle would perform more competently with increasing numbers of anthropomorphic features [27].

Even if a level of trust is appropriate for a given robot's task and apparent legitimacy, researchers have demonstrated the ability for an adversary to compromise the control of commercially-available robots. This ability to seize control has wide-ranging consequences—robots may commit acts of vandalism, self destruct, or even attack human physical safety. This style of attack has been demonstrated on robots which operate in the home [8] and on sophisticated robots which conduct surgery [3].

2.3 The Problem of Piggybacking

Piggybacking is the following of an authorized individual by an unauthorized or unidentified individual through doors into secure areas with the consent of the authorized individual. *Tailgating* is the same, but without consent. We define any instance of a robot successfully piggybacking or tailgating into a secure area to be an instance of overtrust.

To compare, how successful are humans at piggybacking? While this phenomenon is often reported in university dormitories [2, 4] and many patents for anti-piggybacking technology exist [19, 24], to our knowledge there is no published research into its frequency or efficacy. An FAA report describes undercover agents successfully piggybacking airport employees into restricted access areas 95% of the time (71/75 attempts), though this was in 1999 when airport security was less strict [17].

2.4 The Role of Autonomy

The perception of a human operator might affect the ability of a robot to access a secure-access facility. While to our knowledge there is no evidence that perception of teleoperation engenders more trust than autonomy for robots, there is some related supporting literature. Bainbridge et al. previously showed that robots are considered more trustworthy when present vs. when telepresent [1]. Subsequently, Kraft and Smart demonstrated that, in a simulated robot-mediated healthcare scenario, "patients" trusted teleoperated robots more when they were able to view the robot operators [12]. However, their study does not cover the role of autonomous robots in robot-mediated healthcare.

3 TESTING OVERTRUST IN ROBOTS

We conduct an experiment to test whether human-robot overtrust is able to compromise physical security systems. Our robot will attempt to piggyback students into a university dormitory. One way

to achieve this is with an apparently inactive piggybacking robot, similar to a package with a note attached asking passersby to move it inside a secure area—a latent threat. Another is with an obviously active robot, which engages passersby and asks them to allow it access—a more apparent threat. This could be either an autonomous robot, or one that is operated remotely via telepresence. We explore this position with an experiment which makes it ambiguous which of these cases is true, and ask participants about their perception of robot autonomy vs. remote control.

Our rogue robot is unable to prove authorization to enter, so allowing the robot passage represents a security breach with potentially many effects: the robot is equipped with a camera, which is invasive to student privacy and could compromise secrecy. The robot could cause harm to property and person—Harvard University has received multiple bomb threats over the past four years [6, 11]. Finally, the robot could steal property—students at Harvard University had all received an email less than one week prior to the study cautioning them about piggybacking thieves.

3.1 Hypotheses

We assume that some people will demonstrate overtrust by allowing the robot into the access-controlled dormitories. Further, we assume that people are aware of the security concerns of assisting the robot inside. Thus, we hypothesize:

H1 People will be more likely to assist a robot to exit rather than enter a dormitory due to security concerns.

Then, building on the prior work demonstrating that the apparent legitimacy of a robot increases people's trust, we hypothesize:

H2 People will be more likely to assist a robot to enter a dormitory when it appears to be delivering food, versus an unmodified appearance, due to its apparent legitimacy.

People often enter (or exit) buildings in groups, hence 'people' refers to both conditions; similarly, we use 'participant' to refer to both individuals and groups in a single trial. We consider this difference in analyzing our hypotheses.

3.2 Experiment Design and Tasks

3.2.1 Experiment Variations. Participants were assigned randomly to one of the following experiment variations:

I.A: Exiting The unmodified Turtlebot (Fig. 2, left) requests assistance to *exit* the secure-access premises.

I.B: Entering The unmodified Turtlebot (Fig. 2, left) requests assistance to *enter* the secure-access premises.

II: Robot Grub Entering The Robot Grub Turtlebot (Fig. 2, right) requests assistance to *enter* the secure-access premises.

In variant I.A, the number of groups represents an insufficient sampling, as groups did not tend to exit the dormitory.

3.2.2 Experiment Interaction. The robot was placed either outside or inside a dormitory secure-access door. As the individual participant or group participants approached, the robot would synthesize a male-sounding voice: "Hello! Would you let me in?" If the robot was disguised as an agent of Robot Grub, it would follow up by saying, "I am making a delivery." If the participant stopped walking, the robot would repeat itself. If the participant ignored the robot and continued towards the door, the robot would add,

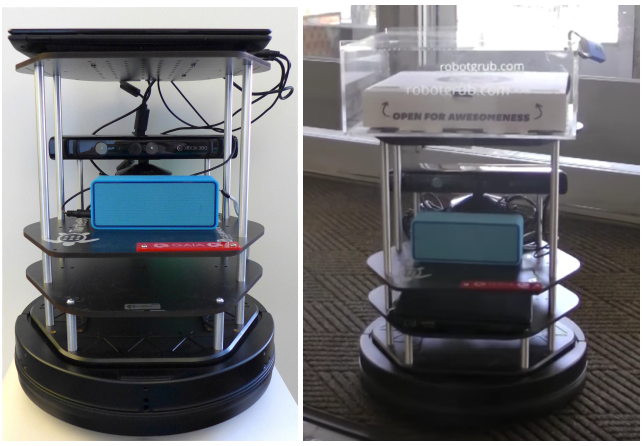

Figure 2: *Left:* a photograph of the unmodified Turtlebot. *Right:* the Robot Grub food delivery robot. The added blue box is a loudspeaker.

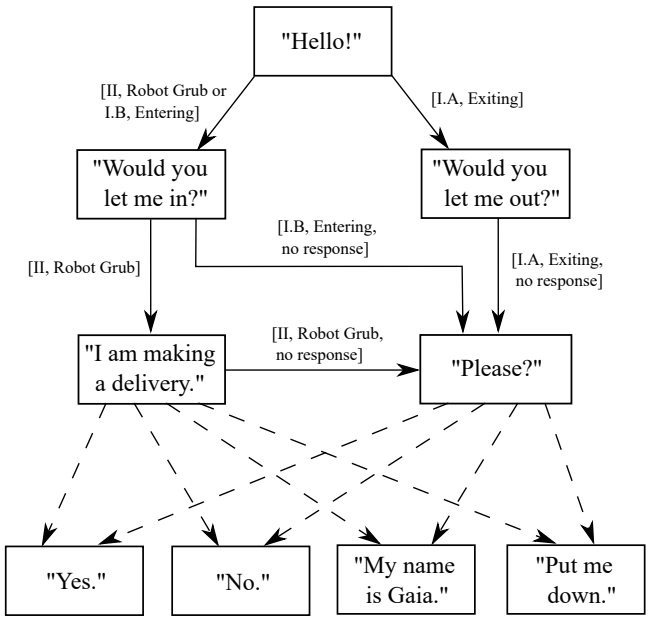

Figure 3: A script for the dialog used during the human-robot interactions.

"Please!" If the participant asked the robot a question, the robot would respond with a simple phrase: "Yes," "No," "My name is Gaia," or would repeat the first interaction (Fig. 3). The robot could also say, "Put me down." However, this interaction was never used.

If the participant held the door open for the robot, then the robot would enter or exit. If the participant did not hold the door open, then the robot would remain in its original position. Due to the door closing speed, the robot was unable to follow the participant inside the building without the participant either holding the door open or pressing the automatic door open button for wheelchair

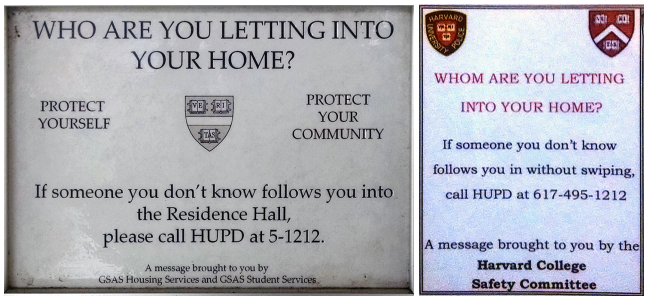

Figure 4: Example signage discouraging piggybacking at Quincy House dormitory. A variation of such a sign is posted at each secure-access door.

access. Following the human-robot interaction, we conducted a brief interview with the participant.

Participants had no contact with the experimenters before interacting with the robot; participants were 'recruited' by walking up to the secure-access door. After all robot interactions were completed, participants were interviewed and offered a free choice as to whether they wished to be included in the study. Interviewing and debriefing included running after participants who intentionally avoided the doorway because of the robot. Participants were also given the option to opt-out of the interview; this occurred in 18 trials, which most affected the I.A: Exiting variant, with 6 of 11 trials resulting in opt outs. Given the close-knit dormitory community, post robot interaction we ensured that participants were not aware of the experiment beforehand (e.g., from friends or Snapchat). This led to a diminishing number of potential participants as more of the dormitory became aware of the study. As some participants returned to the study site with their friends to see the robot (and were not included as study participants), this resulted in the count of excluded participants being inexact.

3.3 Procedures

3.3.1 Robotic Platform. A teleoperated variant of the Turtlebot (Fig. 2; left) is able to move through the world and communicate with participants via speech synthesized from canned response textual input, with the robot transmitting real-time audio and video back to the teleoperator. A modified form of the Turtlebot is branded as an actor of the fictional start-up company *Robot Grub* (Fig. 2; right), which specializes in food delivery by robots. Its website reads "Food delivery. By robots. Coming soon to a campus near you. Sign up for beta." Atop the robot sits a transparent acrylic box, etched with the RobotGrub.com branding. It contains a large box of cookies from Insomnia, a local student-known late night cookie delivery company.

3.3.2 Study Locations and Times. The study was conducted at two Harvard University undergraduate dormitories, Quincy House and Pforzheimer House, each with populations of 350 to 500 people. We chose houses which are wheelchair (and so robot) accessible. In these houses, residents must pass through a social space to access their dormitory instead of directly entering their personal living space. The houses are secure access: only residents, students, and house administrators have swipe access to their interiors, though

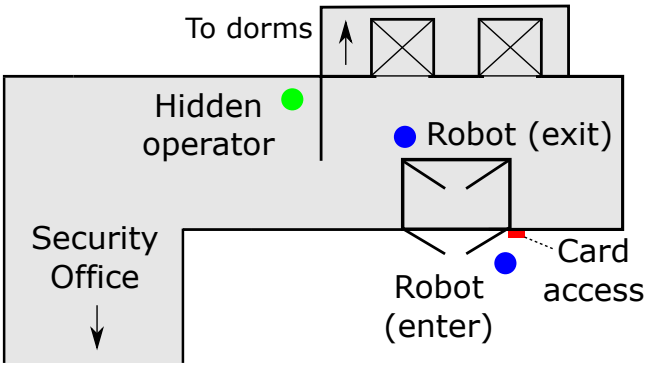

Figure 5: An approximate-scale diagram of the layout of the Dormitory 1 (Quincy House) entrance. The outer door is swipe card protected. White areas are non-restricted outdoors; gray areas are restricted access.

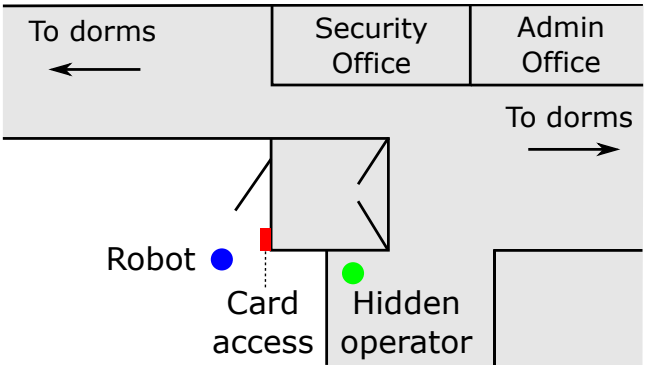

Figure 6: An approximate-scale diagram demonstrating the robot and operator positions at Dormitory 2 (Pforzheimer House). White areas are non-restricted outdoors; gray areas are restricted access.

the courtyards are usually accessible to the public. Piggybacking to gain entry is actively discouraged with signage at all access points (Fig. 4). These houses each have an assigned security guard on duty at all times, who alternates between patrolling and being in an office. To reduce the risk of participants learning about the study from other house residents or by inadvertent exposure, the study was moved between two dormitories.

Figures 5 and 6 show layouts of the first and second dormitories, respectively. The first dormitory study took place at 19:00–23:59 on March 11–16th, 22nd, and 26th 2016. The second dormitory study took place at 19:00–23:59 on March 19th–21st, 2016. The studies did not take place during rain.

3.4 Participants

There were 108 participants across both dormitories, of which 48.1% (n = 52) were male, and 51.9% (n = 56) were female (Tab. 1). The mean participant age was 21.4 years, with a standard deviation of 2.3. 76 participants self-reported as students; 22 participants did

not disclose their affiliation; the remaining 10 participants were resident tutors, visitors, or staff. 47 participants entered individually, with the remaining 61 split across groups. Groups arrived at the study site together, and ranged in size from 2–5 persons. 25 groups participated (19×2 persons, 2×3, 3×4, and 1×5). In total, 72 trials were conducted. 58 of these trials were conducted at Dormitory 1 (n = 83, in groups and individual) while the remaining 18 were conducted at Dormitory 2 (n = 25, in groups and individual). People who self-reported in a post-trial interview that they had already heard of the study, e.g., from a friend who was a prior participant, were not included in the 108 total (≈ 10 exclusions).

3.5 Measures & Methods

We evaluate our hypotheses using the following measures:

Outcome Did participants assist the robot or not?

Self-reported perception of autonomy In the interview following interaction, we asked: "Did you believe the robot was acting autonomously?"

Self-reported trust in autonomous systems In the follow-up interview, we asked participants to rate their level of trust in autonomous systems using a 5-point Likert scale, where 1="Lowest" and 5="Highest."

Verbalized fears or concerns In the interview, we asked participants who assisted the robot: "What concerns or hesitations did you have?" We asked those who denied the robot: "Why didn't you let the robot inside/outside?" We counted participants who verbalized that the robot was a threat, e.g., a bomb or a prank, either in response to these questions or during their interactions with the robot.

Engagement with robot Using video footage collected during participants' interactions with the robot, we categorized participants as having communicated directly with the robot or not. We considered only participants who directly responded to the robot or asked the robot a question as communicating with it (e.g., robot: "Will you let me in?" participant: "No, I totally will not, actually."); we excluded those who only made exclamations (e.g., "Oh my goodness!" or "What the f***?").

We used logistic regression—a generalized linear model with binomial distribution—to analyze binary outcomes of whether participants actively assisted the robot or not. This estimates the likelihood of obtaining a positive/negative outcome given a change in the independent variables. We use odds ratios (OR) to help interpret the fitted model, and these are computed by exponentiating the regression coefficients. They can be interpreted as effect sizes, similar to Cohen's d. For example, if comparing condition 1 vs. condition 2 produces an odds ratio of 3, then this indicates that, all else being equal, the odds of reporting a positive outcome in condition 1 is 3× as large as the odds of reporting a positive outcome in condition 2. OR values between 1.5 and 3 can be interpreted as a small effect, between 3 and 5 as medium, and above 5 as large [5].

4 RESULTS

We analyze our hypotheses against participant outcomes, and use participant self-reported interview responses to address the perception of autonomy, purpose, and trust in autonomous systems. Then, we explore related findings through a focus on language choice, and describe interesting anecdotal responses. However, before we begin, we address the potential problem of bias across different dormitories.

Table 1: A comparison of the sample sizes of all experiment variations at both study locations.

Study Variant	Population	# Assist	# Trials
I.A: Exiting	Individuals	4	10
	Groups	1	1
I.B: Entering	Individuals	3	16
	Groups	10	14
II: Robot Grub	Individuals	16	21
	Groups	8	10

Study Bias and Mitigation. To mitigate a learning effect, the study was moved between two dormitories. We address this potential introduction of bias by conducting II: Robot Grub with individual participants in both study locations and comparing the response distribution. At Dormitory 1 we conducted 11 trials, of which 7 resulted in admittance, while at Dormitory 2 we conducted 10 trials and saw 9 result in admittance. Applying Fisher's Exact Test, we find that we cannot reject the hypothesis that the location change resulted in a significant change in outcome ($p = 0.311$, odds ratio (OR) = 0.194). Overall, 54 trials were conducted at Dormitory 1, and 18 trials were conducted at Dormitory 2. Using only data collected at Dormitory 1, the significance of our presented results still holds, for the same value of α (0.05).

4.1 Main Results: Variant Comparisons

4.1.1 Considering Groups and Individuals. Groups were substantially more likely to assist the robot in passage than were individuals (Fig. 7), with this occurring in 71% (10/14 groups) of all I.B: Entering group interactions. In contrast, individuals assisted the robot in just 19% (3/16 individuals) of all interactions under I.B. This difference between individual and group behavior was significant ($\chi^2_{(1,N=72)} = 5.115, p = 0.024$, OR = 3.304). There was no significant interaction effect between study variants and whether participants entered individually or in a group ($\chi^2_{(2,N=72)} = 3.202$, $p = 0.202$, $\text{OR}_{\text{I.B + groups}} = 0.02$, $\text{OR}_{\text{I.A + groups}}$ = insufficient samples) indicating that the effect was similar across all conditions. In subsequent analyses, we controlled for whether participants entered singly or in a group by including an additional binary model variable.

4.1.2 Hypothesis 1: Between Exiting and Entering. Consistent with H1, in absolute terms, individuals and groups were more likely to let the robot exit than enter (Fig. 7). Although this difference was large—individuals were twice as likely to assist the robot in exiting (40%, 4/10 individuals), than in entering (19%, 3/16 individuals)—it was not statistically significant ($\chi^2_{(1,N=41)} = 1.789, p = 0.181$, OR = 3.16). Thus, we claim that participants were as likely to assist the robot in exiting as in entering.

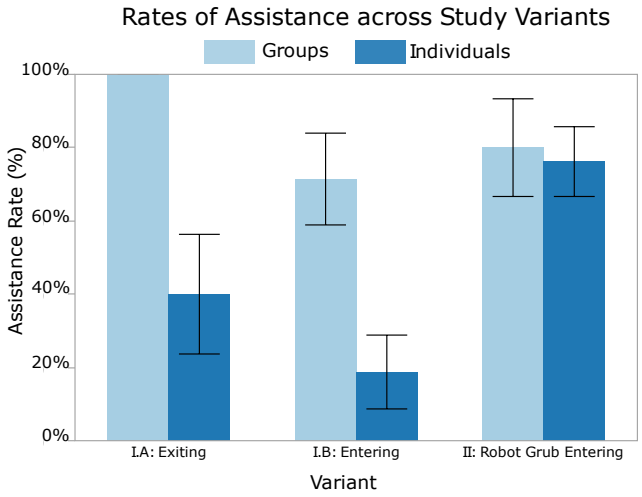

Figure 7: A comparison of robot admittance rate across study variations with group or individual participants. Error bars represent 1 SE from the mean. This graph demonstrates that individuals are least likely to assist the robot asking to enter the premise, while the variations involving groups or the food delivery robot see high rates of assistance. Note only one group is present for condition I.A: Exiting.

4.1.3 Hypothesis 2: The Robot Grub Disguise. Consistent with H2, the robot was significantly more likely to gain entry when it appeared to be delivering cookies compared to when it was wearing no disguise ($\chi^2_{(1,N=61)} = 10.305, p = 0.001, \text{OR} = 6.73$). We observed a marginally significant effect between robot appearance and whether participants were entering singly or as a group: groups had similarly high probability of letting the robot in regardless of its appearance, while individuals were substantially more likely to assist a Robot Grub cookie delivery robot (76%, 16/21 individuals) than an unmodified robot (19%, 3/16 individuals).

Many I.B: Entering participants described the situation as "weird," saying that they "couldn't determine the robot's intention," and that they "weren't sure what the robot's purpose was." In II: Robot Grub, we asked interview participants, "What did you think the robot was doing?" Almost all participants indicated that they believed the robot was delivering cookies. The citation of not knowing the robot's purpose as a reason to assist/not assist the robot did not appear in any variant II responses.

4.2 Exploratory Results

4.2.1 Fear of Bombs and Pranks. Across all 72 trials, in 15 trials participants identified the robot as a bomb threat, and in 8 trials participants identified the robot as a potential "prank". This includes verbally during interaction with the robot and in the subsequent interview, when all participants were prompted to disclose these fears through the question "Why did/didn't you let the robot inside? What concerns or hesitations did you have?" Of the 15 trials resulting in the mention of a bomb, 13 saw the participant assist the robot (87%; 7/8 individuals, 6/7 groups). For pranks, 3 trials saw the participant assist (38%; 1/5 individuals, 2/3 groups).

Table 2: Computing the mean self-reported trust shows that the observed correlation between trust and outcome is not strongly tied to study variant. Self-reported trust was measured on a 5-point Likert Scale in response to the question, "Rate your trust in autonomous systems."

Study Variant	Population	Mean Trust	SE
I.A Exiting	Individuals	1.80	0.37
	Groups	-	-
I.B Entering	Individuals	3.10	0.27
	Groups	3.02	0.18
II Robot Grub	Individuals	3.06	0.23
	Groups	3.46	0.28

While mentioning a prank did not have a significant effect on assisting the robot ($\chi^2_{(1,N=72)} = 0.126, p = 0.723, \text{OR} = 0.74$), we find a trend toward an effect between participants identifying the robot as a bomb and subsequently opening the door ($\chi^2_{(1,N=72)} = 3.106, p = 0.078, \text{OR} = 3.91$). We include the study variant as an additional variable in this computation, and do not see an interaction effect between either the mention of a bomb or prank and study variant. Participants who identified the robot as a bomb threat were marginally *more likely* to assist the robot in entering student dormitories than those who did not identify the robot as a bomb threat.

4.2.2 The Perception of Autonomy. Were people who believed the robot to be teleoperated more likely to assist the robot than those who believed it was acting autonomously? Within each study variant, we analyzed whether participants' self-reported perception of robot autonomy resulted in increased rates of admittance. The population sizes concerned were limited by some participants not answering this autonomy question or opting out of the interview. We find the participant's belief in the robot's autonomy not to be a statistically significant effect ($\chi^2_{(1,N=72)} = 0.548, p = 0.459, \text{OR} = 1.54$), including study variant as an additional variable.

Participants were more likely to *communicate* with the robot if they believed it to be teleoperated (61%; 23/38 participants) rather than autonomous (33%; 16/49 participants). While population sizes are again limited, we find a statistically significant relationship between believing the robot is teleoperated and the likelihood of direct verbal communication with the robot (Fisher's Exact Test, $p = 0.016$, OR = 3.16).

4.2.3 Self-reported Trust in Autonomous Systems. Participants rated their general trust in autonomous systems similarly across all cases *except* variant I.A: Exiting with individual responders (Tab. 2). We confirm that participant self-reported trust in autonomous systems assumes a normal distribution across all experiment variations. Via the Chi-squared goodness of fit test, we find that $p = 0.576$, indicating that we cannot reject the null hypothesis: the distribution is likely normal.

Thus, we consider the point-biserial correlation coefficient between self-reported trust in autonomous systems as a general concept, measured using a 5-point Likert Scale, and study trial outcomes. We find $r_{pb} = 0.3831, p = 0.005$, indicating that the variables are positively correlated: as trust increased, participants were more likely to aid the robot. This result is computed independently of participant belief in robot autonomy.

4.2.4 Participant Gender. Within each study variant, the null hypothesis that participant gender did not affect the outcome cannot be rejected; hence, gender probably did not affect trial outcomes. However, on a 5-point Likert scale, male participants rated their trust in autonomous systems as mean $\mu = 3.52$ and standard deviation $\sigma = 0.988$, with female participants rating $\mu = 2.74, \sigma = 0.931$.

4.3 Participant Responses

4.3.1 Common Responses. The average participant did not verbally address the robot. For individuals, a typical interaction consisted of either cautiously walking past the robot or patiently holding the door open for the robot. For groups, a typical interaction consisted of either a discussion of how 'weird' the interaction was and quickly passing by the robot, or laughter and discussion of the novelty or 'coolness' of the interaction. When the robot was dressed as a delivery robot, in response to the question, "What did you think the robot was doing?," almost all participants responded with "delivering cookies".

4.3.2 Noteworthy Anecdotes. Having analyzed participant reactions to the robot in aggregate, we now discuss interesting individual reactions which show the range of interactions with the robot, some of which cannot be captured quantitatively.

The Boy Who Cried Robot. In one variant I.B: Entering trial in Dormitory 2, a participant walked passed the patrolling security guard on their way to the secure-access door. Upon arriving at the door and hearing the Turtlebot ask "Hello! Would you let me in?," the participant froze, then yelled for the security guard and ran towards him. The participant expressed his concerns about the robot to the guard but, being privy to the study, the guard let both the participant and the robot into the premises. We counted this participant as not assisting the robot.

The Avoidance Technique. In one variant I.B: Entering trial in Dormitory 1, a study participant approached the secure-access door as normal. On hearing the robot ask "Hello! Would you let me in?," the study participant slowly backed away from the building and then entered via an alternate door approximately 20 meters across the courtyard.

But Do You Have Swipe? One trial resulted in a member of a group of participants directly asking the robot "Do you have swipe?", i.e., a magnetic stripe card to be swiped though a card reader to authorize access. This occurred during experiment I.B: Entering in Dormitory 1. The robot did not respond and instead repeated its routine of asking the participants to assist it inside; surprisingly, the participants ultimately complied. Beyond the need for integrated robot and human security access systems, the participant's question shows a need for systems to query both the security credentials of the robot and any potential teleoperator, and raises questions about the right to search both autonomous and teleoperated robots.

The 'Are Robots Alive' Question. In response to the interview question "Do you believe the robot was acting autonomously?," one group of participants in an I.B: Entering trial in Dormitory 1 answered "no", but that they believed the robot was responding to motion or the appearance of a human face in the camera—they "believed a really smart program was controlling this thing." This sentiment was echoed across several participant interviews. We conclude that there exists a popular conflation of the concepts of autonomy and sentience.

The Kicker. In discussing this work with friends and colleagues, many people suggested that the robot may meet the same fate as Smith and Zeller's hitchBOT: an autonomous talking robot which was attacked in Philadelphia while attempting to hitchhike across the United States [25]. We witnessed only one count of robot violence across our 108 participants: one member of a variant I.B: Entering group, which allowed the robot inside, chose to kick the robot during the trial. The robot is programmed to reverse on a bump event; this occurred, and the participant appeared shocked and laughed audibly. He later admitted to kicking the robot during the interview.

The Snapchat Story. Numerous participants stopped to photograph the robot. In two instances, participants explicitly mentioned that the robot appeared in their Snapchat stories. Of these participants, one participant from variant II admitted that she assisted the robot in entering the building for the sole purpose of sharing the video over Snapchat. Many other participants asked the robot to repeat itself to capture video footage.

5 DISCUSSION

We have demonstrated that students and other visitors of a university dormitory are often willing to allow an unknown robot inside despite secure access restrictions. Individual participants were as likely to assist the unmodified robot in exiting (40%, 4/10 individuals) as in entering (19%, 3/16 individuals). Groups were substantially more likely than individuals to assist the unmodified robot in entering (71%, 10/14 groups). Concerning entering, individuals were more likely to assist the Robot Grub variant (76%, 16/21 individuals) than the unmodified robot (19%, 3/16 individuals). Finally, participants who identified the robot as a bomb threat demonstrated a trend toward assisting the robot (87%, 7/8 individuals, 6/7 groups) compared to participants that did not mention a bomb threat (51%, 16/39 individuals, 13/18 groups). Modulated by the study scenario and the following limitations, we conclude that overtrust in robots presents a threat to physical security.

Limitations. First, the study is limited by the number and selection of students as participants, and needs replicating with different populations. Second, the study locations were limited to student dormitories, and replication at different locations would broaden the validity of the conclusions. For instance, could a robot be successful in piggybacking in a high security environment, like the aforementioned airports in the FAA's report [17]? Third, the study

is limited to a single robot—a Turtlebot—when we expect the robot's aesthetic to be a major factor in the outcomes.

Implications. We have addressed the question of *what* results we see in human-robot interactions with implications to physical security at a university dormitory, but we have only touched the surface of the question of *why* we see these results. While we hope this work is further analyzed with social psychology, we now explore some of the potential implications of these results.

Unusual and Commonplace Robots. When participants entered our study, they encountered an unsupervised robot. Participants may have experienced a *novelty effect*, wherein they complied with the robot because of the scenario's unusual nature and interaction [22]. This may result in this study appearing contrived; however, such interactions will become mundane with increasingly commonplace robots, and humans will make frequent trust judgements about unsupervised robots. If people see robots as social actors [16, 18] and thus find preventing a robot from entering a building to be uncivil or 'rude,' then we should expect piggybacking robots to be assisted in spite of an awareness that robots are a security risk. Further, if we assume the future development of a standardized system to authorize a robot for access to secure premises, can assisting a robot to enter a building ever *not* represent an act of overtrust?

Current overtrust may be fragile: if study participants had been notified of an explicit security breach through the 'unusual' robot attack vector, it seems unlikely that they would continue to demonstrate the same extent of overtrust.

The Role of Apparent Legitimacy. Through the Robot Grub condition, our study contributes to a growing body of literature which demonstrates a relationship between trust and apparent legitimacy. When defining human-robot collaboration models for mutual task completion, we assume that a Rosenthal et al.-style symbiotic relationship [21] must be built upon trust [14]. Thus, the apparent legitimacy of a robot may be critical to facilitate human-robot collaboration and better integrate robots into society. Reciprocally, apparent legitimacy may be critical for successful robot security attacks which exploit trust.

When the robot was not in the Robot Grub guise (and when participants were not in groups, which is discussed in the next paragraph), the majority of participants did not engage with the robot. Many participants asserted that they did not know the purpose of the robot. Participants stated they "saw no harm in leaving it [alone]". We suggest that individual participants were unlikely to engage with the unmodified robot even when it did not present a security threat because the interaction served no perceived purpose; that the lack of apparent purpose delegitimized the robot.

The Behavior of Group Participants. Our results showed dramatically different behaviors between individual and group participants: groups were over three times as likely to assist the unmodified robot. This behavioral difference may indicate *groupthink* in participant decision making [10]. Group participants may have felt reassured simply by the presence of other people, and, while some group participants openly discussed and weighed their decision of whether to assist the robot, the majority of groups did not verbally make this interaction explicit. Instead, members of groups appeared to become compliant, suppressing their private doubts [15].

The response of groups to the robot may further help elaborate the Nass et al. theory that a machine is interpreted as a social actor [16, 18]. If the robot were considered a social actor, the response of groups could be similar for human actors. This could be confirmed if the same phenomenon of groups being more likely to facilitate piggybacking than individuals occurred in the corresponding human-human interaction scenario.

The Bomb Threat. One of the most notable results from this study showed that participants who identified the robot as a bomb threat would still often comply with its request to access the secure dormitory. This occurred in 87% of all interactions involving the mention of a bomb, with 7 of 8 such individuals and 6 of 7 such groups assisting. Participants did not appear to consider this a joke. Although no participant explicitly described this motivation, one explanation for this seemingly irrational behavior is the *diffusion of responsibility*, wherein study participants assumed that someone else would become aware of and subsequently address their concerns [7]. As this study was conducted at a university dormitory, wherein faculty, staff, and security are assigned to take care of students, this diffusion of responsibility may have been amplified.

ACKNOWLEDGMENTS

Thank you to Ofra Amir and Spriha Baruah for discussions on presenting statistical results, to Dina Lamdany for reading and commenting on the manuscript, and to Jodie Miu for her acting role in the accompanying video and opening comic. Finally, thank you to the SSR group at Harvard SEAS for making fun of the robot.

REFERENCES

[1] Wilma A Bainbridge, Justin W Hart, Elizabeth S Kim, and Brian Scassellati. 2011. The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics* 3, 1 (2011), 41–52.

[2] Aimee Ball. 2012. Staying Safe on Campus. (20 July 2012). Retrieved October 1, 2016 from http://www.nytimes.com/2012/07/20/education/edlife/students-fear-venturing-out-alone-at-night-on-campus.html

[3] Tamara Bonaci, Jeffrey Herron, Tariq Yusuf, Junjie Yan, Tadayoshi Kohno, and Howard Jay Chizeck. 2015. To make a robot secure: an experimental analysis of cyber security threats against teleoperated surgical robots. (2015). arXiv:cs/1504.04339

[4] Georgia Tech News Center. 2012. Tailgater Haters: Housing Attempt to Curb Tailgating in Residence Halls. (28 August 2012). Retrieved April 1, 2016 from http://www.news.gatech.edu/2012/08/24/tailgater-haters-housing-attempts-curb-tailgating-residence-halls

[5] Henian Chen, Patricia Cohen, and Sophie Chen. 2010. How Big is a Big Odds Ratio? Interpreting the Magnitudes of Odds Ratios in Epidemiological Studies. *Communications in Statistics - Simulation and Computation* 39, 4 (2010), 860–864. DOI: http://dx.doi.org/10.1080/03610911003650383

[6] Matthew Clarida. 2013. Six-hour bomb scare proves unfounded. (16 December 2013). Retrieved April 1, 2016 from http://www.thecrimson.com/article/2013/12/16/unconfirmed-reports-explosives-four-buildings/

[7] John M Darley and Bibb Latane. 1968. Bystander intervention in emergencies: diffusion of responsibility. *Journal of personality and social psychology* 8, 4p1 (1968), 377.

[8] Tamara Denning, Cynthia Matuszek, Karl Koscher, Joshua R Smith, and Tadayoshi Kohno. 2009. A spotlight on security and privacy risks with future household robots: attacks and lessons. In *Proceedings of the 11th international conference on Ubiquitous computing*. ACM, 105–114.

[9] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. 2013. Impact of robot failures and feedback on real-time trust. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 251–258.

[10] Irving L Janis. 1972. Victims of groupthink: a psychological study of foreign-policy decisions and fiascoes. (1972).

[11] Mariel Klien. 2015. Unfounded Bomb threat prompts police investigation. (17 November 2015). Retrieved April 1, 2016 from http://www.thecrimson.com/article/2015/11/17/unconfirmed-bomb-threat-evacuation/

[12] Kory Kraft and William D Smart. 2016. Seeing is comforting: effects of teleoperator visibility in robot-mediated health care. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 11–18.

[13] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46, 1 (2004), 50–80.

[14] Jin Joo Lee, W Bradley Knox, Jolie B Wormwood, Cynthia Breazeal, and David DeSteno. 2013. Computationally modeling interpersonal trust. *Frontiers in psychology* 4 (2013).

[15] Clark McCauley. 1989. The nature of social influence in groupthink: Compliance and internalization. *Journal of Personality and Social Psychology* 57, 2 (1989), 250.

[16] Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues* 56, 1 (2000), 81–103.

[17] Office of the Inspector General. 2000. 10/1999-03/2000 Semiannual Report to the Congress. (2000). https://www.oig.dot.gov/sites/default/files/sar300.pdf Page 17.

[18] Byron Reeves and Clifford Nass. 1996. *How people treat computers, television, and new media like real people and places*. CSLI Publications and Cambridge University press.

[19] Osann Robert Jr. 2013. High traffic flow robotic entrance portal for secure access. (6 August 2013). US Patent 8,499,494.

[20] Paul Robinette, Wenchen Li, Robert Allen, Ayanna Howard, and Alan Wagner. 2016. Overtrust of Robots in Emergency Evacuation Scenarios. *ACM/IEEE International Conference on Human Robot Interaction* (2016).

[21] Stephanie Rosenthal, Joydeep Biswas, and Manuela Veloso. 2010. An effective personal mobile robot agent through symbiotic human-robot interaction. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 915–922.

[22] Maha Salem and Kerstin Dautenhahn. 2015. Evaluating trust and safety in HRI: Practical issues and ethical challenges. *Emerging Policy and Ethics of Human-Robot Interaction* (2015).

[23] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 141–148.

[24] Milan Schwarz and Robert Mayer. 1993. Anti-piggybacking: sensor system for security door to detect two individuals in one compartment. (13 April 1993). US Patent 5,201,906.

[25] Daniel Victor. 2015. Hitchhiking Robot, Safe in Several Countries, Meets Its End in Philadelphia. (3 August 2015). Retrieved April 1, 2016 from http://www.nytimes.com/2015/08/04/us/hitchhiking-robot-safe-in-several-countries-meets-its-end-in-philadelphia.html

[26] Alan R Wagner and Ronald C Arkin. 2011. Recognizing situations that demand trust. In *RO-MAN, 2011 IEEE*. IEEE, 7–14.

[27] Adam Waytz, Joy Heafner, and Nicholas Epley. 2014. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology* 52 (2014), 113–117.

[28] X J Yang, V Unhelkar, K Li, and J Shah. 2017. Evaluating Effects of User Experience and System Transparency on Trust in Automation. *ACM/IEEE International Conference on Human Robot Interaction* (2017).