Research Article Open Access

Daniel J. Brooks, Dalton J. Curtin, James Kuczynski, Joshua J. Rodriguez, Aaron Steinfeld, and Holly A. Yanco*

A Communication Paradigm for Human-Robot Interaction During Robot Failure Scenarios

Abstract: When autonomous robot systems experience failures, communication about the failure to both the people responsible for the robot and to people who happen to be nearby is critically important. New robot users as well as bystanders might not be familiar enough with a system to tell whether a robot is working properly and experienced robot operators might not notice signs of trouble due to being out-of-the-loop. Thus, an important feature for robots will be the ability to communicate failure to humans when failures occur. In this paper, we describe a study conducted with a smartphone-based feedback system we designed to explore push and pull forms of communication. We found the communication methods improved participants' understanding of robots' state, increased their confidence in interacting with the robots, and allowed them to remotely monitor and control the robots.

Keywords: Human-robot interaction, situation awareness, bystander interaction, failure recovery

Daniel J. Brooks: Toyota Research Institute, E-mail: dan@tri.global
Dalton J. Curtin: University of Massachusetts Lowell, E-mail: daltoncurtin@gmail.com
James Kuczynski: University of Massachusetts Lowell, E-mail: jkuczyns@cs.uml.edu
Joshua J. Rodriguez: University of Massachusetts Lowell, E-mail: Joshua_Rodriguez@student.uml.edu
Aaron Steinfeld: Carnegie Mellon University, E-mail: steinfeld@cmu.edu
***Corresponding Author: Holly A. Yanco:** University of Massachusetts Lowell, E-mail: holly@cs.uml.edu

1 Introduction

Robot systems, including self-driving cars, delivery drones, and cleaning robots, are becoming more common in public settings. Soon, people will interact with and live alongside these and other types of autonomous robotic systems on a regular basis. Due to the scale and complexity or these robot systems, we cannot expect every interaction to be flawless even after systems have matured. For example, a robot performing a necessary behavior that is perceived as inexplicable or unpredictable can have a detrimental effect on people's situation awareness and lead to negative user experiences.

Such issues will affect not only the robots' users but also bystanders, who may have marginal awareness of or interest in the robots' capability or mission. Humans will increasingly be in situations requiring them to make decisions about unsupervised and unfamiliar systems, some of which may be critical to their own safety. To make matters worse, there are currently no standards – de facto or otherwise – to serve as a guide for allowing people to communicate with or to influence the behaviors of these machines. Therefore, we believe that robots need efficient and understandable methods for bidirectional communication, even at a basic level, which could be used for robot operators or for people who are bystanders to the robot. As a first step towards a solution to this need, this paper describes our findings in an experiment using push and pull notifications on a smartphone to communicate robot information to minimally trained participants.

2 Related Work

Taxonomies have been developed that categorize faults and provide insight into the many complex ways a system could fail [2, 20] and attributes of "dependable" systems are available [16]. There will likely never be perfectly reliable robots, so strategies are needed for mitigating the consequences of failure. Effective methods from prior work include providing advanced warning or confidence feedback, apologizing after failure, asking for help from bystanders, and failure-specific natural language requests [4, 14, 15, 19].

Unfortunately, there are also examples where recovery strategies produce negative effects. For example, inexplicable robot behavior can lead to misassignment of blame and humans often fail to recognize their own

correctable mistakes [13]. Likewise, robot assignment of blame can lead to very negative reactions and lower trust [5, 12].

Failures can be expressed implicitly if the robot is able to clearly communicate its intentions in a way that can be contrasted with physical behavior when an error occurs. For example, if a drone were equipped with a light ring direction indicator (e.g., [21]) that was indicating a straight flight path while the drone was translating to the side, users familiar with the drone's normal operation would be able to immediately discern that something was not right. Likewise, a robot could point its arm where it believed a person wanted it to go, thereby providing an opportunity to intervene [7].

A valuable failure mitigation strategy for certain situations is to ask for human intervention. Cha et al. [3] summarized the process of asking for help as having three phases: 1) getting someone's attention, 2) indicating to the person that help is needed, and 3) conveying the request for help.

Unfortunately, there appear to be nuances in how and when such requests will be honored. For example, in one experiment, participants in a public kitchen area were asked for assistance with coffee preparation by an approaching robot [8]. Only half of the participants complied with the robot's request. The vast majority of the people who helped were not busy concentrating on another task (one of the experimental conditions), while most who were busy ignored the robot, tricked it into thinking they had given it the coffee so it would go away, or shut the door to keep the robot out.

Deciding who to ask for help is also important. Asking the same person for help frequently could quickly become annoying. Rosenthal et al. [19] addressed this issue by distributing the burden across an office hallway and anticipating who might be available based on prior behavior. After a few days, many people closed their office doors.

There have also been explorations of interaction with openly imperfect robots. For example, Yasuda and Matsumoto [22] hypothesized that people may relate well to imperfect robots, viewing them as similar to children or infants who try but fail in their efforts. Their interaction design led to positive experiences for most participants.

Prior work has also investigated creating introspective systems capable of explaining why a robot behaved in a particular manner by tracing and logging the flow of information through a system, and keeping track of which pieces of data were used in making progressively higher level decisions [1]. However, providing users with information about the cause of a failure could also make the situation worse. For example, Kim and Hinds [13] found that robots that attempt to explain their ambiguous actions and errors can actually decrease people's perceived understanding of the system.

3 Interaction Design

Smartphones have rapidly become a ubiquitous technology based on their value in a wide variety of tasks. Due to their market saturation, we believe these devices to be an ideal proxy through which humans and autonomous systems could communicate with each other.

To explore methods for human-robot communication, we developed a smartphone-based interaction method, with two types of interaction. The first type uses *pull-style interactions* (Figures 1 and 2) to enable people to query information or communicate with the robot. Pull-style interactions waited in the background for the user to initiate interaction, rather than interrupt the user. The second type used *push-style interactions* (Figure 3), where the robot initiated communication with nearby people. Push-style interactions interrupted the user and were thought to be useful when alerting people of hazards or requesting help. These interaction methods were used in conjunction with a balloon popping game that we created for use in the experiment.

We note that robot vacuum manufacturers have been shipping their products with accompanying apps for a few years. For example, Neato Robotics' robot vacuums [18] have an app that provides four push notifications: "Done Cleaning," "Low Battery," "Base Connected to Power," and "Move Base to New Location" [17]. As another example, iRobot's robot vacuums [10] have an app that provides two types of push notifications: "Done Cleaning" and error messages [9]. iRobot's robot vacuum users can also use the "Care" section of the app to see the status of the robot, its bin, and its extractors [11]; these notifications are pull notifications. While these manufacturers have developed apps to be used with their robots, we are not aware of any published studies that investigate how these types of interactions influence human-robot interaction when a robot needs assistance.

For both interaction types in the smartphone app that we designed for our experiment, robots were identified with a unique name and an icon that looked like the robot, in order to help identification of the robot(s); see Figure 4 for the set of robots used in the experiment.

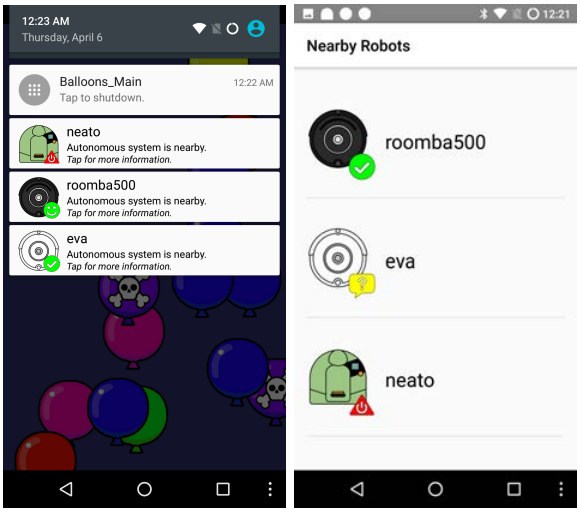


Fig. 1. Pull-style notifications. The left image shows the pull notifications accessed from within the balloon game. The right image shows the pull notifications by opening the smartphone app directly.

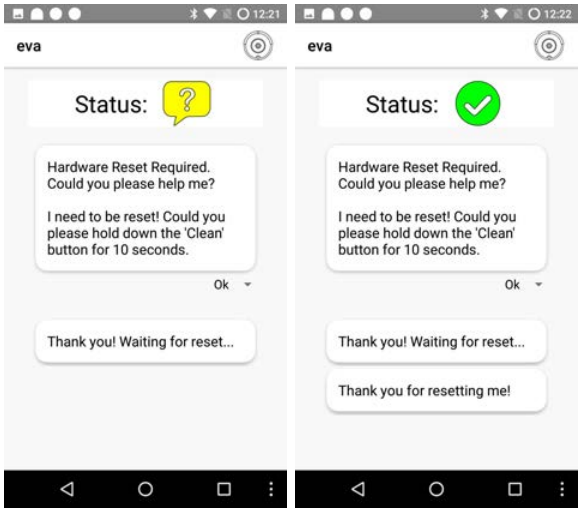


Fig. 2. Acting on pull-style notifications: The lower left shows one robot's help screen reached from the pull notifications within the app and the lower right shows the robot's help screen after the reset.

In the pull-style interactions, the robot's icon also had a small status icon overlaid on its lower right corner, to provide immediate feedback about the robot's state, without needing to open the message.

4 Experiment Methodology

The experiment described in this section represents our initial investigation into the use of smartphones as a platform for establishing a ubiquitous communication paradigm designed to allow people to gain basic information from and interact with a variety of autonomous robots. In this experiment, we used several types of robot vacuum cleaners (Figure 4) as our platforms to provide clearly describable task and to allow interaction with multiple robots in a fairly small space.

This study was approved by the University of Massachusetts Lowell's Institutional Review Board. This research has complied with all relevant national regulations and institutional policies.

We asked participants to perform two simultaneous tasks: (a) manage a fleet of vacuum cleaning robots in "cleaning" an area of a floor while (b) playing a balloon popping video game on a smartphone at the same time. The robots were secretly modified in a way that allowed them to appear as if they were still standard robot vacuum cleaners. They were programmed to intentionally exhibit various problems with their functionality which the participants would need to address, while the video game acted as a distracting secondary activity for diverting user attention from the robots. Our objective was to gain insight into how participants would manage the two tasks and the robots when provided with access to the different communication styles (i.e., push and pull). Since we wanted to understand the impact on novices, participants were not provided with any training on the use of the app; we also limited training on the robots themselves, as described below.

We used a within-subjects design where each participant experienced two runs – one with only the manufacturer's on-robot interfaces (i.e., smartphone communication *disabled*) and one with the manufacturer's robot interfaces supplemented with our communication app (i.e., smartphone communication *enabled*). In both cases, the participants had access to the on-robot manufacturer's button interfaces.

Our participants were divided into two equally sized groups that were counterbalanced. One group started with *disabled* communication in the first run, then had *enabled* communication in the second run; the other group started with *enabled* communication in the first run, then had *disabled* communication in the second run.

Participants were asked to perform two simultaneous tasks in each of their runs: use a set of three robots to collect plastic beads scattered around on the floor and play a simple video game on a smartphone. The video game task was a simple, skill-less game of balloon popping. Animated balloons drifted up from the bottom of the screen to the top and "popped" when touched, earn-

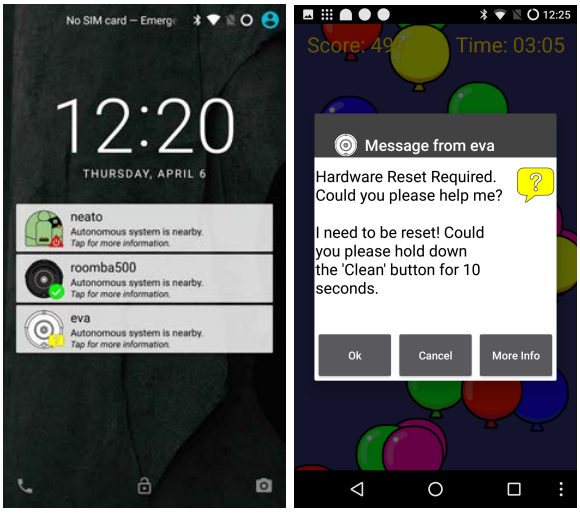


Fig. 3. Design for push notifications. The left image shows push notifications on the smartphone's lock screen. The right image shows a push notification from a robot over the balloon game.

Fig. 4. The set of robots used in the experiment. Neato (far left) and the Discovery model Roomba (far right) were disabled to include "broken" robots in the experiment design.

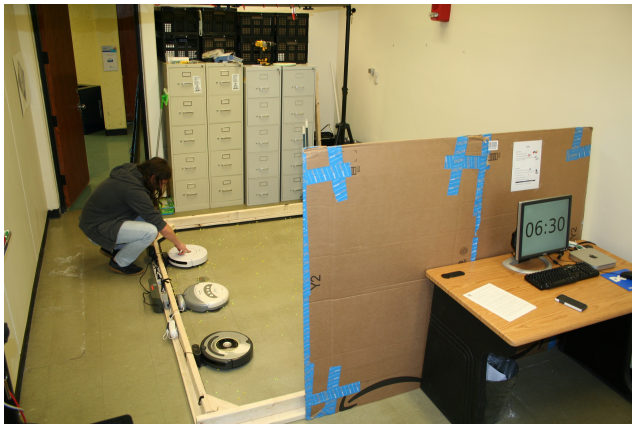


Fig. 5. One of the authors demonstrating what a participant might do in the cleaning zone to assist a robot.

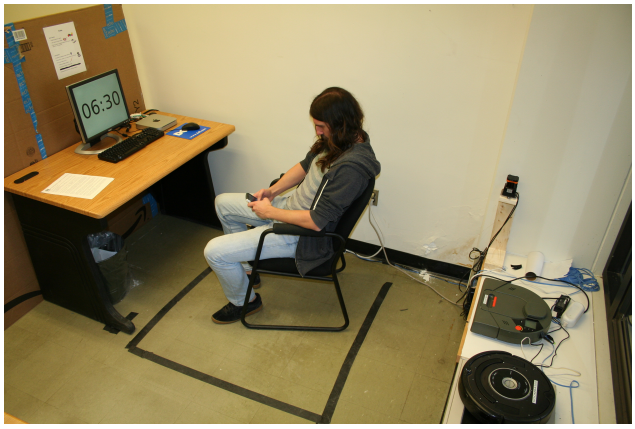


Fig. 6. One of the authors demonstrating what a participant would be doing if in the game playing zone.

ing the participant points. Participants needed to pay attention, as balloons marked with skull and crossbones symbols would take away earned points if popped.

In parallel, participants were asked to use three apparently standard robot vacuums, which were actually modified and programmed to exhibit the desired behaviors, to collect small plastic beads scattered around the floor inside an area fenced off with a low wood frame. Participants were required to remain outside of this area and were not permitted to collect the beads themselves, forcing them to use the robots to accomplish the task. Participants could interact with the robots at the edge of the cleaning zone, as shown in Figure 5.

In order to track time spent physically interacting with or observing the robots, participants were only allowed to play the game when inside a marked game-playing zone that did not have a clear view of the robots (see Figure 6). The game interface was disabled when exiting this zone, thereby preventing gaining or losing points until back in the zone.

Participants were not briefed or trained on how symbols on the robot's hardware or beeps the robots communicated to the user would corresponded to what was causing the robots to be unable to operate, nor were participants showed how to start the robots. They were shown all the robots that would be used during the experiment including the robots that were to be swapped in on the second run and how to empty the dust bins on each robot. Similarly, participants were not briefed or trained on the use of the smartphone app for robot notifications.

Participants were given 6.5 minutes to both play the balloons game and use the robots to collect beads. A digital clock in the game-playing zone and a timer in the

balloons game provided awareness of this period. Participants were responsible for keeping track of the time remaining by using either a digital clock showing the game time that was positioned inside the game-playing zone, or by using the game timer inside the balloons game (which also showed the experiment time).

While participants earned money based on points scored in the game, the total was adjusted by the fraction corresponding to the percentage of beads collected by the robots (i.e. if they collected 70% of the beads they would get to keep 70% of the points scored in the game). This percentage was determined by measuring the weight of the collected beads, as compared to the known amount scattered on the floor at the start of the run (roughly 100g). At the end of each run, the beads collected by the robots were measured by weight to calculate the percentage collected, while the remaining beads were removed from the cleaning area by the experimenter. Participants would lose an additional 100 points (approximately 45 seconds worth of balloons game playtime) for each robot that was not on a charging station when time ran out. Participants received compensation based on the higher of their two final scores. Each person received $5 for simply completing the study, and could earn up to an additional $10 based on their performance, using a score based on the time that the two working robots spent running and the balloons game score.

In addition to an informed consent form before the experiment, participants were asked fill out a pre-experiment questionnaire, two post-run questionnaires, and a post-experiment questionnaires.

4.1 Hypotheses

Our hypotheses in this study were as follows:

Hypothesis 1 (H1): Participants would be able to determine which robot they were communicating with using our system, despite similarities between robots.

Hypothesis 2 (H2): Participants would be able to retrieve information about the robots they were working with, identify solutions to problems faster, and allocate their time more appropriately when using the smartphone-based interface compared to only having the default manufacturer interfaces on the robots.

Hypothesis 3 (H3): Participants would prefer having access to the additional information provided by our smartphone-based interface over using the default manufacturer interface on the robot alone.

	Phone		Robot Starting Conditions				
Cond	Support	Run	R_A	R_B	R_C	R_D	R_E
1	Enabled	1	Easy	Dead	Help		
1	Disabled	2	Help			Easy	Dead
2	Enabled	1	Easy			Help	Dead
2	Disabled	2	Help	Dead	Easy		
3	Disabled	1	Easy	Dead	Help		
3	Enabled	2	Help			Easy	Dead
4	Disabled	1	Easy			Help	Dead
4	Enabled	2	Help	Dead	Easy		

Table 1. Experiment conditions, as described in Section 4.2

4.2 Independent Variables

This experiment used a within-subjects (repeated measures) design in which each participant performed two runs. While both runs included the manufacturers' default interfaces on the robots themselves, our smartphone-based communication was only enabled in one. Between runs, the experimenter replaced two of the three robots used in the previous run in full view of the participant as part of "resetting the task," while the participant filled out the post-run questionnaire. (The two robots not being used in a particular run were on charging stations behind the participant in the game playing zone, as shown in Figure 6.)

Conditions were counterbalanced by assigning participants into one of four experimental conditions corresponding to the two independent variables in the experiment: the order in which the smartphone app support was used in the runs and the order in which two different starting configurations were used.

The experiment used 5 robot vacuum cleaners: four iRobot Roombas and one Neato XV11. The four Roombas were two working Roomba 500 models (R_A and R_D), a working Roomba 600 model (R_C), and a non-working Roomba Discovery model (R_B). The Neato XV11 (R_E) was intentionally programmed to not work. Three of these vacuums were used during the first run, after which two of the three were replaced before starting the next run. There were two combinations of robots that were switched between, each of which consists of two "working" robots and one "non-working" robot: *Group 1* consisted of R_A, R_B and R_C while *Group 2* consisted of R_A, R_D, and R_E.

Each of the three robots exhibited a different level of functionality: one robot was "easy" to start, simply requiring the push of a button; one required "help" from the participant before it would start running (it needed to be reset); and the last one played "dead" and would never start working (Table 1). The two "working" robots performed their default cleaning behaviors except that the length of time they ran for was shortened and some "problems" were introduced. Two minutes after a robot started cleaning, it would automatically start returning to its dock. After the "easy" robot returned, it would require the participant to come to empty its dustbin before it would be able to start cleaning again. In contrast, the "help" robot could immediately be told to resume cleaning (even before it finished returning to its dock) when its two minutes were up, and never required the dustbin to be emptied.

The robot that needed to be reset flashed a red LED ring around the power button, illuminated a red error symbol in the shape of a circle with an exclamation mark in the center of it, and would periodically play a distinct error tone until the participant pressed and held down on the power button. When the smartphone app was enabled, this robot sent a push-style interaction, causing a message to appear on the smartphone. The reset event occurred at the beginning of each run.

The robot that needed its dustbin emptied illuminated a yellow LED ring around the power button, flashed a blue LED labeled "dirt detect" and would occasionally play an different error tone until the dustbin was removed. When the smartphone app was enabled, participants could view information about the full dustbin on the robot's status page (a pull-style interaction).

Time spent trying to make the third robot work was wasted. One of the dead robots had no lights on and showed no indication it even had power; the other dead robot had a single lit LED light, but showed no other signs of being functional. The dead robot could be viewed using pull-style interactions in the smartphone app, which would reveal the robot's broken status.

4.3 Dependent Variables

Aside from the three questionnaires, data was logged on the robots and smartphone, in manual notes, and video recordings. After each run, we documented the resting positions of the robots, the percentage of beads collected, and the game score. The post-run questionnaire included the NASA's Task Load Index (TLX) Questionnaire [6], as well as questions about participant confidence in the app and where participants thought app information originated.

Logged data included the time in game, the number of times the game was started and stopped, when and how often they left the game zone, and the game score over time. App specific data was also captured, including the number of times each robot's app page was accessed via a pull-style interaction, the time spent viewing each robot's app page, the number of times background notifications were observed, and whether the participant explicitly agreed or declined (via the app dialog buttons) to help the robot.

5 Results

Twenty people (14 men, 6 women) between ages 18 and 33 participated, all of whom had previous experience with smartphones. Two people had prior experience using Roombas and one person had used a Philips robot vacuum.

The majority of the analyses were 2x2 mixed-groups factorial ANOVAs on app status (enabled, disabled) and the participants' assigned starting scenario.

5.1 Robot Usage

The app was a significant main effect for the combined time the robots spent cleaning [$F(1,16) = 17.052, p < 0.001$] (Figure 7). There was a significant interaction between the experiment condition and the presence of the app [$F(3,16) = 3.49, p = 0.04$], however there was no main effect for condition by itself [$F(3,16) = 1.064, p = 0.39$] (Figure 8). A post-hoc two-tailed t-test showed that the time the robots spent cleaning was significantly higher during the run in which people had access to the app ($M = 406, SD = 131$) compared to the run in which it not provided ($M = 307, SD = 98$); $t(19) = 3.49, p = 0.002$. In other words, participants were able to keep the robots cleaning for more of the session when using the app. While this was especially true for people who used the app during the second run, it was also generally true for people who used the app in the first run.

5.2 Game-Playing Zone

People spent less time outside the game-playing zone, and thus less time watching and physically interacting with the robots, when they used the app. The

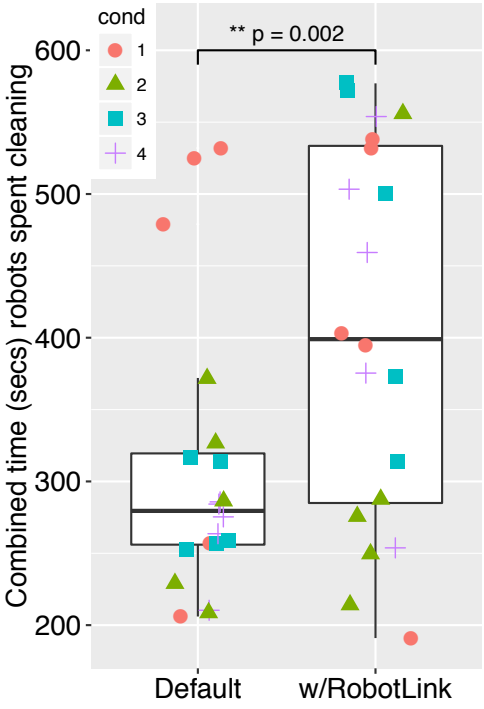


Fig. 7. Time robots spent cleaning with an without the app.

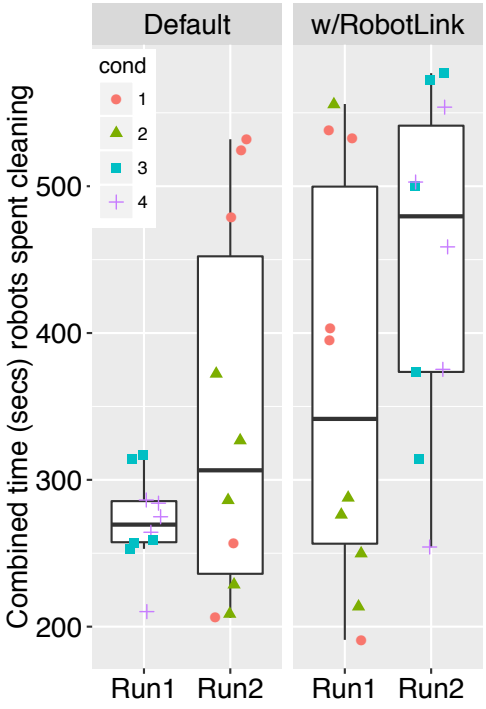


Fig. 8. Time robots spent cleaning by run number.

app was a significant main effect on the amount of time participants spent outside the game-playing zone $[F(1,16) = 4.589, p = 0.048]$ (Figure 9). A post-hoc two-tailed t-test showed that the time participants spent outside the game-playing zone observing and interacting with robots was significantly less during the run in which they had access to the app (M=151, SD=53) compared to the run in which it was not provided ($M = 176$, $SD = 44$);$t(19) = -2.23, p = 0.038$.

Participants context switched between tasks less often when provided the app, presumably because it allowed them to determine if the robots needed attention without exiting the game-playing zone. The app was a significant main effect on the number of times participants switched between being inside and outside the game-playing zone $[F(1,16) = 4.47, p = 0.05]$ (Figure 10). A post-hoc two-tailed t-test showed that the number of switches was significantly less during the run when they had access to the app ($M = 4.1, SD = 1.92$) compared to the default ($M = 5.1, SD = 2.57$); $t(19) = -2.078, p = 0.05$.

5.3 Balloons Game

On average, participants had lower scores on the balloons game when they were also using the app. Presence of the app was a significant main effect on the balloons game score $[F(1,16) = 7.567, p = 0.01]$. A post-hoc two-tailed t-test showed that participants scored significantly fewer points during the run in which they had access to the app ($M = 382, SD = 137$) compared to the defaul ($M = 463, SD = 131$); $t(19) = -2.66, p = 0.01$ (Figure 11).

The lower scores are probably due to a decrease in time spent playing the balloons game, since the rate at which players scored points was similar between conditions, and there was no discernible difference in the number of penalties incurred. This aligns with the significant main effect for the app on the amount of time spent playing the balloons game $[F(1,16) = 4.52, p = 0.05]$. A post-hoc two-tailed t-test showed that participants spent significantly less time playing the balloons game during the run in which they had access to the app ($M = 180, SD = 33$), compared to the default ($M = 206, SD = 56$); $t(19) = -2.21, p = 0.04$ (Figure 12).

5.4 Robot Interactions

There was a weak main effect of having access to the app on the number of times participants pressed buttons on the robots $[F(1,16) = 3.70, p = 0.07]$. There was no main effect of the experiment condition $[F(3,16) =$

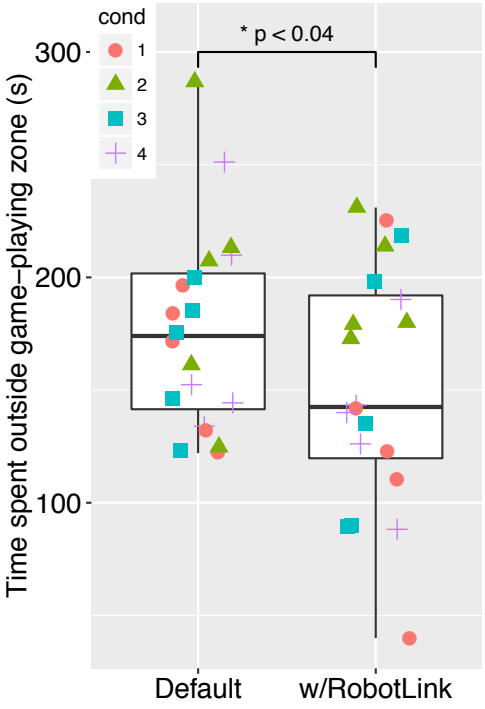


Fig. 9. Time spent outside of the game-playing zone without and with the robot status app.

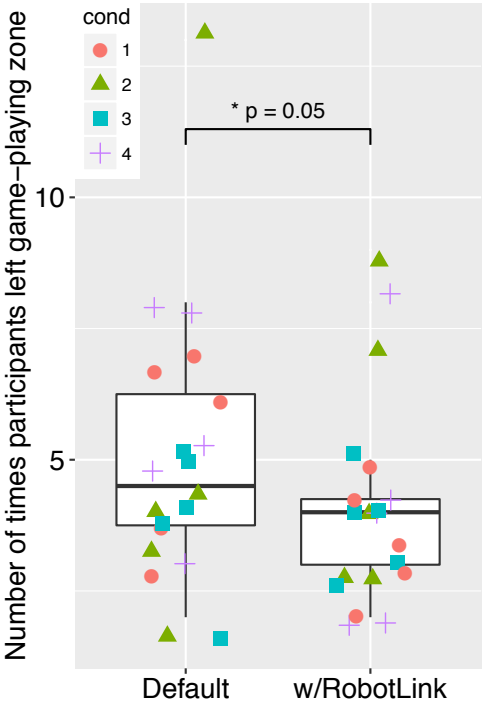


Fig. 10. Number of switches in and out of the game-playing zone without and with the robot status app.

$1.79, p = 0.18$] and there was no significant interaction between the app and the experiment condition [$F(3, 16) = 1.04, p = 0.4$] on the number of times participants pressed buttons on the robots. A post-hoc two-tailed t-test showed (weak significance) that the number of times participants pressed buttons on the robots was fewer during the run in which they had access to the app ($M = 10.85, SD = 9.6$) compared to the run in which it was disabled ($M = 15.7, SD = 6.56$); $t(19) = -1.92, p = 0.07$ (see Figure 14). Simply put, people spent less time using the robots' physical interfaces when they also had access to the app. Two potential explanations for this are that people were using the app controls instead of the physical controls, and that they potentially had a better understanding of why a robot might not be responding to their actions.

As implied above, app availability led to significant differences in participants' physical interactions with robots (Figure 15). Time with robots was calculated by coding each interaction's start and stop times for every robot up until the time the robot that needed help was reset. Timing started whenever the participant knelt or leaned over a robot while reaching towards, touching, or looking at the robot or the phone. Timing was stopped whenever the user stood up, moved their hand away from the robot, or looked away from the robot or the phone. Time spent working with the dustbins was excluded, with time stopping when the dustbin was removed and re-starting once it was replaced. Inter-rater reliability of two coders (one experimenter and one researcher not involved with data collection, both included on the IRB protocol) was computed using Cohen's Kappa and showed significant agreement ($\kappa = 0.87, \alpha = 0.05$).

A pairwise comparison using a two-tailed paired t-test on the time spent with robots during the default run without the app showed a significant difference ($p = 0.03$) between the time spent with the robot that needed to be reset ($M = 28, SD = 27$) and with the robot that was playing dead ($M = 14.8, SD = 19$). The difference between the robot which needed its dustbin emptied ($M = 24.3, SD = 20$) and the robot that needed to be reset was not significant ($p = 0.6$), nor was there a significant difference between the dustbin robot and the robot playing dead ($p = 0.1$). In comparison, a pairwise comparison using a two-tailed paired t-test of the time spent with robots during the run with the app showed a larger significant difference ($p = 0.008$) between the time spent with the robot which needed to be reset ($M = 21.5, SD = 22$) and the robot that was playing dead ($M = 4.6, SD = 10.5$). There was also a significant difference between the robot which needed

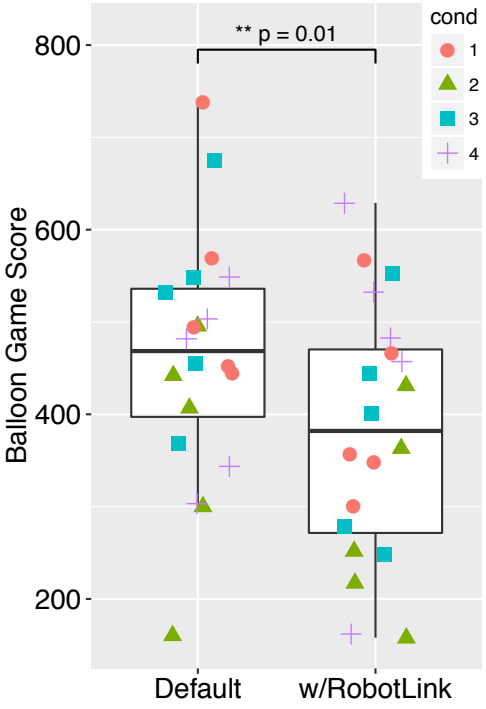


Fig. 11. Score in the balloons game, without and with the robot status app.

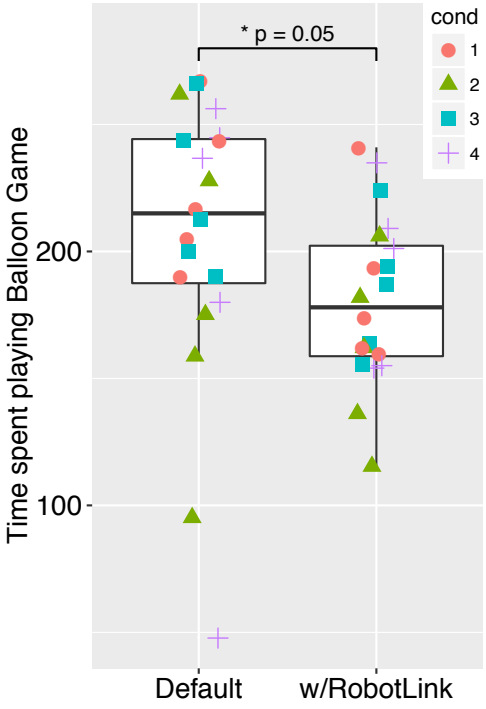


Fig. 12. Balloons game playing time, without and with the robot status app.

to be reset and the robot that needed its dustbin emptied ($M = 6, SD = 13.5, p = 0.02$), but not between the robot which needed its dustbin emptied and the robot that needed to be reset ($p = 0.5$). This suggests the app helped participants more efficiently understand which robot they needed to reset.

Participants are required to hold down the clean button for 7 seconds in order to reset a robot. This would enable a robot that needed to be reset to continue cleaning. 7 seconds was chosen to reduce the chance of a user from accidentally resetting the robot. No participant who started without the app held down the clean button to reset the robot that needed help in the first run for longer than 5 seconds (Figure 16). People with the app in their first run appeared to apply their experiences to their second run without the app, as shown by the steadily increasing number of times people held down the button for over 5 seconds in the second run. This same group may have also either been uncertain about which robot they needed to reset, or guessed the same technique would work on multiple robots since they also appear to have tried to reset the "dustbin" robot more than in any other situation. Everyone who had the app in their second run successfully reset the robot that needed help, and very few attempts were made at trying to reset a robot that had not requested help.

Half of the participants (10/20) viewed all three robots' status pages using the app. The robot that needed to be reset used a push-style interaction which caused a popup message to appear on the smartphone screen, interrupting what the user was doing. As a result, each of the participants ended up viewing this robot's page at least once. The robot that needed its dustbin to be emptied was viewed by 12 of the 20 participants, and the robot that played dead was viewed by 10 of the 20 participants.

Half of the participants were able to use the app to view information about 2 or more robots. Using the data from the 10 participants who viewed all three robots' status pages, a pairwise comparison using a two-tailed paired t-test showed significant differences on the time participants spent on those app pages between the robot that needed to be reset ($M = 88.7, SD = 43.7$), needed to have its dustbin emptied ($M = 45.3, SD = 17.1$), or was non-functioning ($M = 9, SD = 9$) (Figure 17). Participants spent significantly more time looking at the page of the robot that needed to be reset than the robot that needed its dustbin emptied ($p = 0.02$) or the robot that had been disabled ($p < 0.001$). They also spent significantly more time looking at the page of the robot whose

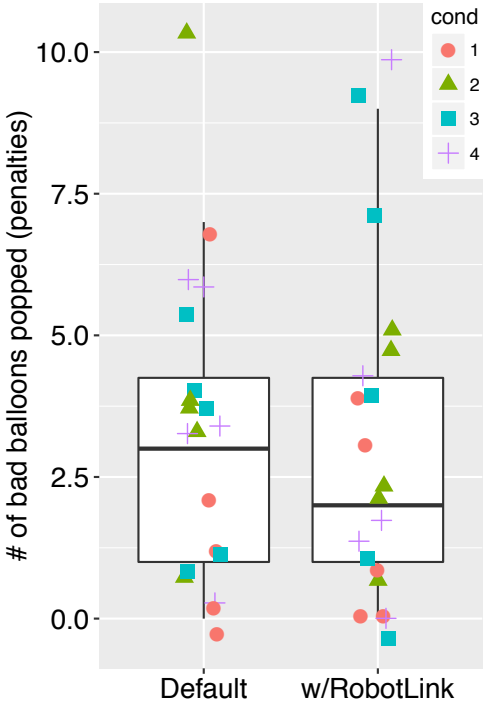


Fig. 13. Penalties incurred in the balloons game, without and with the robot status app.

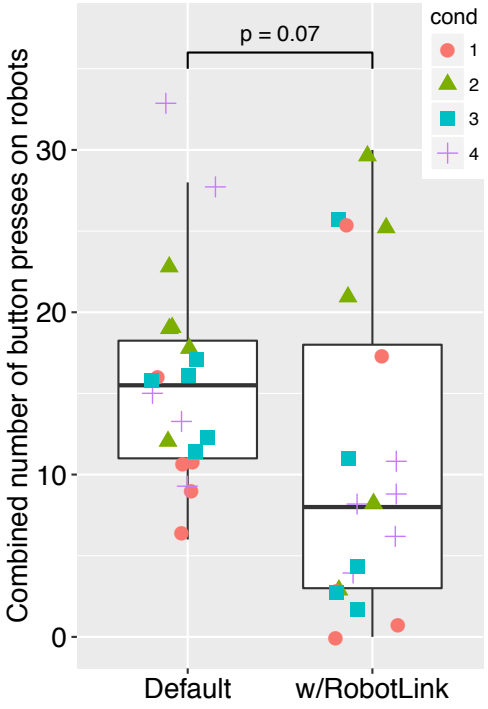


Fig. 14. Number of button presses on the robots, without and with the app

dustbin needed to be emptied than that of the robot which had been disabled ($p < 0.001$). In other words, the amount of time people spent viewing information about the different robots was associated with the appropriate amount of attention needed to get and keep each robot working.

5.5 Post-Run Questionnaires

After each run, participants were asked to fill out a questionnaire asking about their experience with the robots, their perception of what help (if any) they needed to provide to robots, and their workload.

In each of the two runs, the three robots each engaged in one of three distinct behaviors. One robot immediately required being "reset" before it could begin cleaning, but once this had been completed could continue running without needing any other help. A second robot was immediately available to begin cleaning upon request, but thereafter needed to periodically have its dustbin emptied before it could continue cleaning. The last robot played dead, and simply refused to work for the entire duration of the run. During both runs, the two robots that needed help would emit visual (flashing lights) and auditory indicators (beeping sounds) to signal there was a problem until the issue was resolved. Following each run, participants were asked *"What kind of help did the robot(s) require or request? Select all that apply."*. With a single exception, participants' responses were limited to the two actions which they actually needed to take. More people correctly identified the two solicited actions during the second run (24) than during the first run (19).

The number of participants who understood a robot needed to be reset was significantly higher ($p = 0.04$ using McNemar's test) during runs when the app was present (12/20) compared to the default (5/20). The run number did not significantly effect participants' understanding of whether or not a robot needed to be reset ($p = 0.5$). However, the order in which the phone was used did seem to have an effect; 5/10 participants who had access to the app during the first run understood they needed to reset one of the robots, compared to 7/10 who had the app during the second run. In comparison, without the app only 2/10 people during the first run and 3/10 from the second run understood that one of the robots needed to be reset.

Participants generally felt that the robots were predictable (Figure 18a) regardless of whether or not they had access to the app (no significant difference was

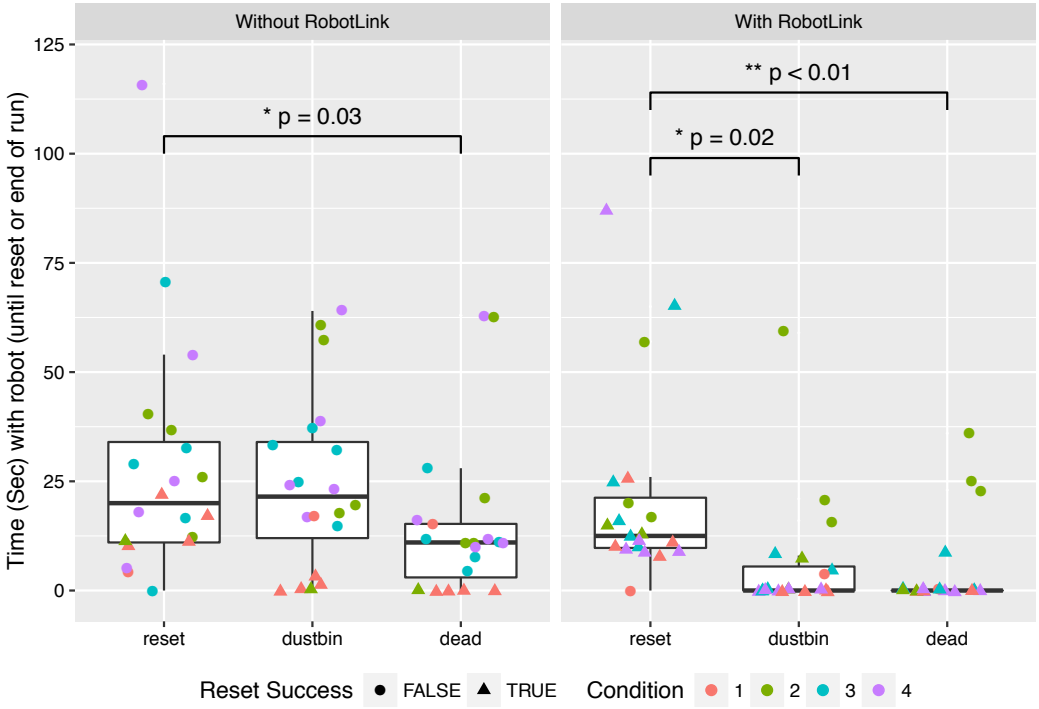


Fig. 15. Time physically spent with robots until robot reset, without and with the robot status app.

found using a two-tailed paired t-test; $t(19) = -0.31, p = 0.7$).

People were more confident that they understood what robots were doing while using the app. A two-tailed paired t-test showed that people reported significantly more confidence in their understanding of the robots' behaviors during runs in which they had the app ($M = 4.65, SD = 1.57$) compared to runs in which they did not ($M = 3.7, SD = 1.42$); $t(19) = 3.13, p = 0.005$ (Figure 18b).

People's satisfaction with the robots was higher (weak significance) during the runs when the app was enabled ($M = 4.6, SD = 1.19$) than during runs in which it was not ($M = 4.05, SD = 1.23$) according to a two-tailed paired t-test; $t(19) = 1.93, p = 0.07$ (Figure 18c).

A two-tailed paired t-test also showed that participants found it significantly easier to determine what was needed to keep each robot working during when they had the app ($M = 5, SD = 1.26$), compared to the default ($M = 3.7, SD = 1.75$); $t(19) = 3.21, p = 0.004$ (Figure 18d).

As part of each post-run questionnaire, participants completed the NASA TLX questionnaire. Six 2x2 mixed-groups factorial ANOVA were performed to examine the effects of using the app and the experiment conditions on mental and physical workload, perceived performance and success, and how how rushed and discouraged they felt.

There was a weak significant main effect of having access to the app on mental workload [$F(1, 16) = 4.26, p = 0.055$]. Also, there was a significant main effect of the experiment condition on mental workload [$F(3, 16) = 3.3, p = 0.04$], and a weakly significant interaction between the app and the experiment condition on mental workload [$F(3, 16) = 3.02, p = 0.06$]. A post-hoc two-tailed paired t-test showed (with weak significance) that participants tended to have a lower mental workload during the run in which they had the app ($M = 3.7, SD = 1.13$) than the run without it ($M = 4.1, SD = 1.33$); $t(19) = -1.8, p = 0.088$. A pairwise two-tailed t-test with Holm correction that was used to compare differences in mental workload between experiment conditions found significant differences between condition 1 ($M = 4.8, SD = 1.14$) and conditions 3 ($M = 3.4, SD = 0.97, p = 0.048$) and 4 ($M = 3.1, SD = 0.074, p = 0.045$), but not between any of the other experiment conditions.

These findings suggest that using the app lowered mental workload, but that perception of workload was influenced by prior experience. We found that people who were placed in Conditions 3 and 4 had lower mental workloads than people in Condition 1. This is interesting because people used the app during their second run

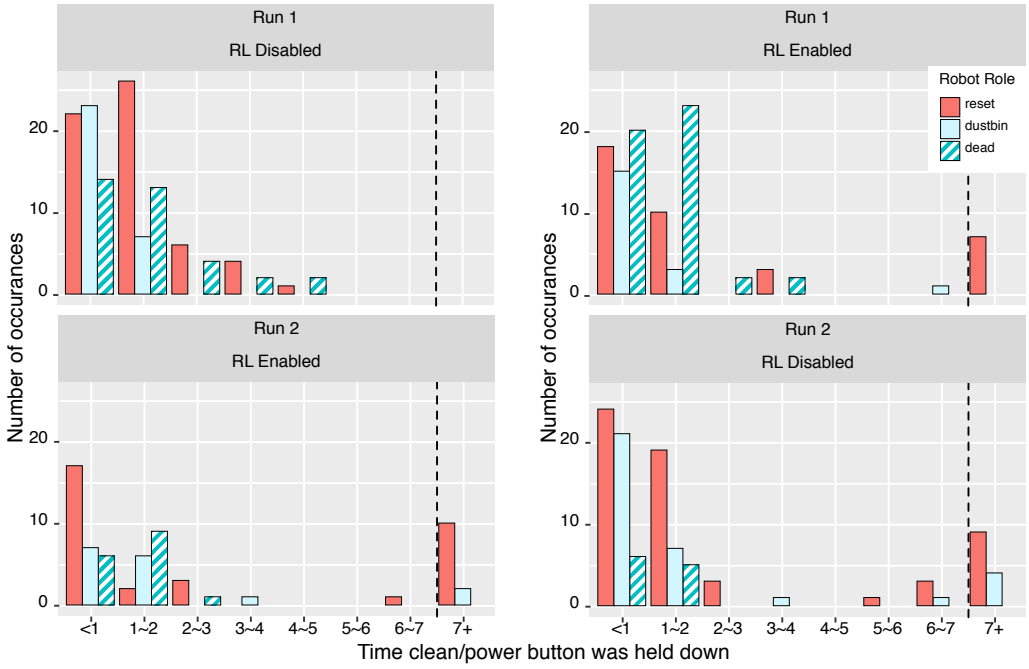


Fig. 16. Clean button click/hold times (in seconds)

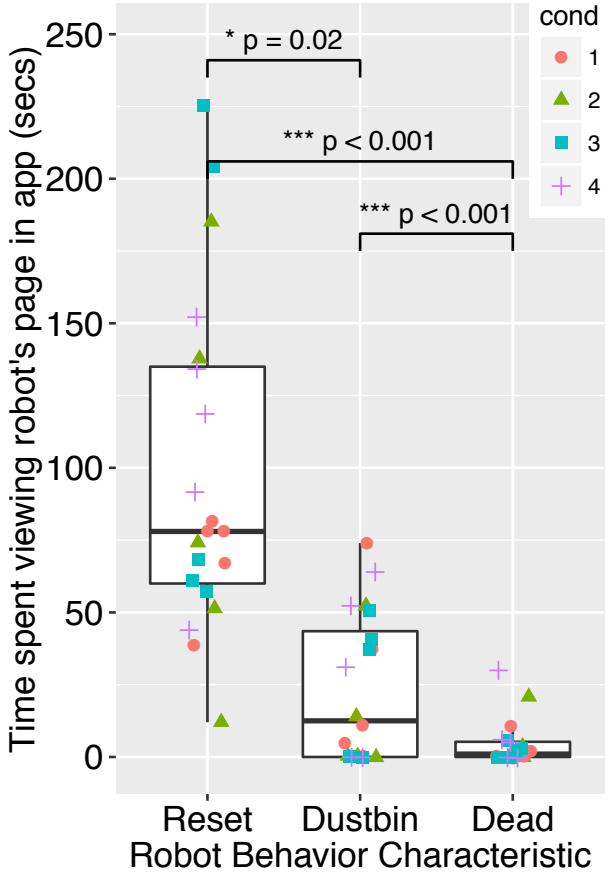
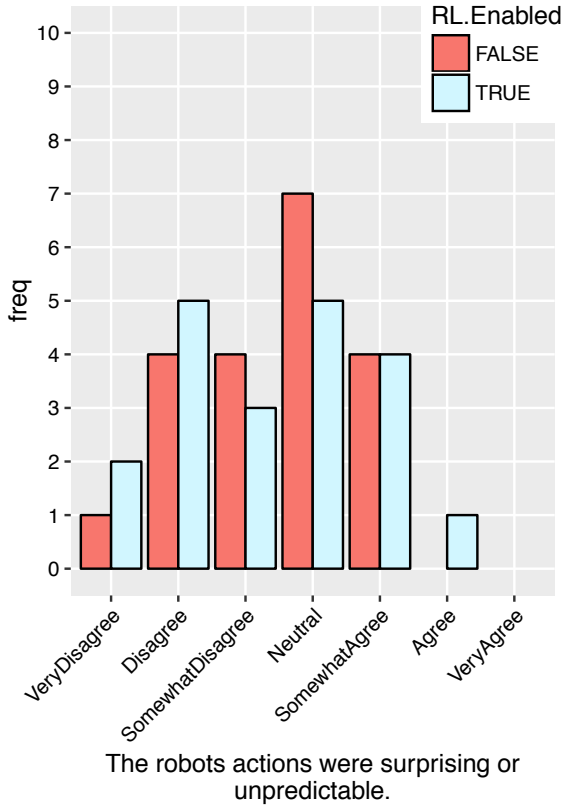


Fig. 17. Time spent in the robot status smartphone app

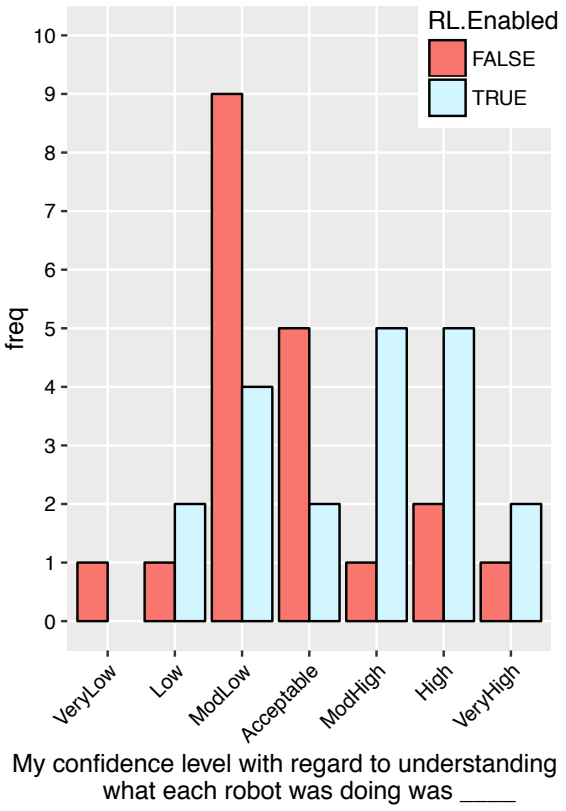
for both Conditions 3 and 4, while people in Condition 1 used the app in the first run (the same was true of Condition 2, but the difference was not as pronounced as in Condition 1.)

There was a significant main effect of having access to the app on the physical workload $[F(1,16) = 4.65, p = 0.046]$; however, there was no main effect for experiment condition $[F(3,16) = 1.80, p = 0.18]$ on physical workload, nor was there a significant interaction between the variables $[F(3,16) = 1.06, p = 0.39]$. A post-hoc two-tailed paired t-test showed that participants tended to have a lower physical workload during with the app $(M = 2.15, SD = 0.81)$ compared to the default $(M = 2.7, SD = 1.34)$; $t(19) = -2.15, p = 0.04$. The app allowed people to substitute walking between different locations in the room to accessing robots with an interaction in the app.

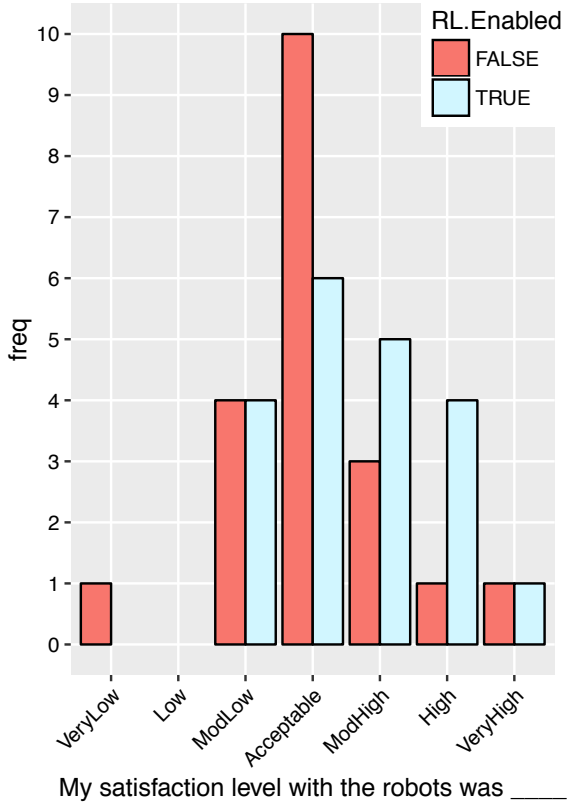
There was a significant main effect of having access to the app on how discouraged participants reported feeling $[F(1,16) = 5.04, p = 0.04]$; however there was no main effect of the experiment condition $[F(3,16) = 1.54, p = 0.24]$, nor was there a significant interaction between the variables $[F(3,16) = 0.34, p = 0.79]$. A post-hoc two-tailed paired t-test showed that participants reported feeling less discouraged during when they had the app $(M = 3.4, SD = 1.67)$ than without it $(M = 4.05, SD = 1.62)$; $t(19) = -2.37, p = 0.03$. This is consistent with our findings that people were more



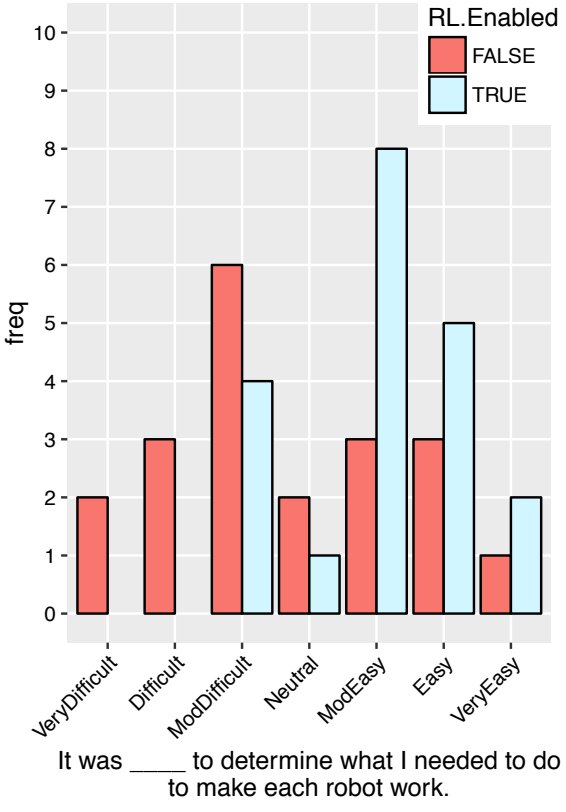
(a) Robot predictablity



(b) Confidence



(c) Satisfaction



(d) Obtaining Information

Fig. 18. Participant robot reviews, by app usage

confident about understanding the robots' actions and found it easier to determine what to do to make robots work while using the app.

5.6 Post-Experiment Questionnaires

After the second run, participants were asked to complete a final questionnaire. Nearly all of the participants (18/20) reported that learning how to use the app was easy (Figure 19a). Most also reported that it was easier to figure out what they needed to do with the robots (14/20) and to control them (15/20) by using the app rather than by physically looking at the robots themselves (Figures 19b and 19c).

All 20 participants said they preferred having access to the app (Figure 20a). Nineteen out of the 20 participants thought the information they received from the app was helpful and informative, and 16/20 felt more in control of the robots when they had access to the app (Figures 20b and 20c). There were no significant differences between run ordering conditions ($p = 0.3, p = 0.3$, and $p = 0.4$, respectively, using two-tailed t-tests).

5.7 Experiment Scores

There was no main effect for the app on experiment score $[F(1,16) = 1.16, p = 0.3]$. However, there was a significant main effect for experiment condition on experiment score $[F(3,16) = 3.18, p = 0.05]$, and there was a weak interaction between the app and experiment $[F(3,16) = 3.06, p = 0.06]$. A post-hoc pairwise comparison using paired t-tests showed significant differences between Condition 2 ($M = 196, SD = 75$) and Conditions 1 ($M = 184, SD = 92, p = 0.03$) and 3 ($M = 183, SD = 78, p = 0.04$).

We have been unable to produce a suitable explanation for why participants in Condition 2 did not perform as well as those in the other three conditions. The low experiment scores were a combination of both low balloon game scores (although the differences were not significant) and low combined time robots spent cleaning. The later was the result of just a single person (out of five) in Condition 2 successfully using more than one robot, compared to 4 out of 5 in Condition 1, 5 out of 5 in Condition 3, and 4 out of 5 in Condition 4. There were no significant differences between Condition 2 and the other conditions with respect to time spent outside the game-playing zone, app usage, time spent playing the balloons game, penalties incurred in the balloons game, or time spent physically interacting with the robots. That said, despite the lack of difference in experiment scores based on the presence of the app, our results still largely supported our hypotheses.

6 Discussion

6.1 H1 is Supported by the Results

A majority of people indicated they could tell which robot was the source of information in the app (70%) and that it was easier to figure out what they needed to do to make the robots work (70%) with the app than by looking at the robots. Six of the ten participants who had the app during their first run were able to "reset" the robot that needed help, and five out of ten were able to transfer that knowledge, by successfully identifying and resetting a different robot that needed help in their second run without using the app.

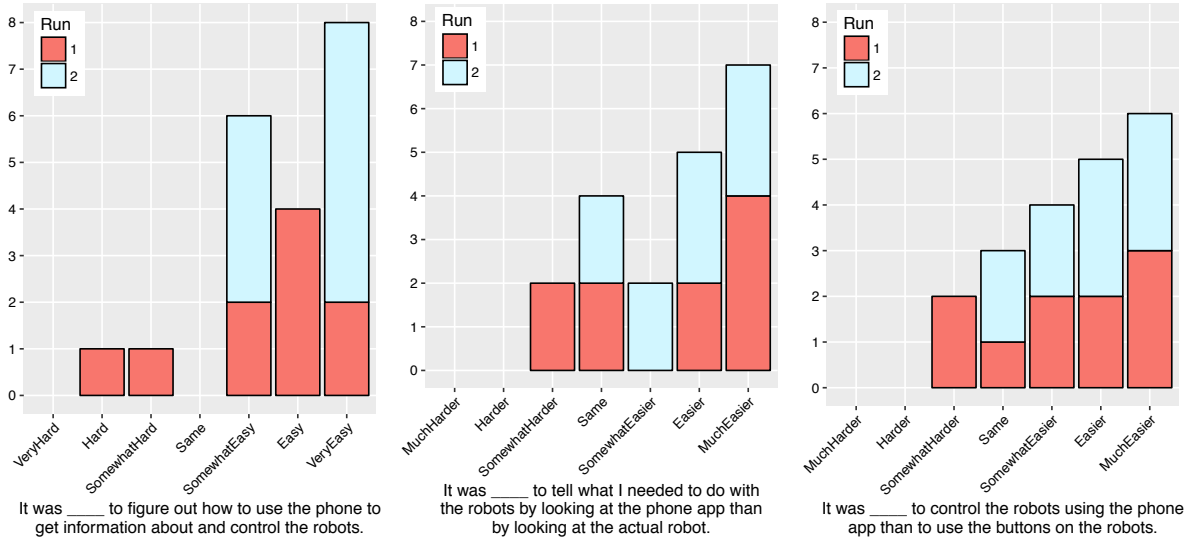
In contrast, no one who did not have the app in the first run was able to reset the robot that needed help. However, *all of those same people* (10/10) were able to successfully reset another robot which needed help when provided the app in their second run.

6.2 H2 is Partially Supported by the Results

Participants were able to use the app to retrieve information about the robots and to identify solutions to problems. All of the participants used the app to view the robot that employed a push-style interaction (popup message), and half of the participants viewed all three robots using the app. There are a few possible reasons why more of participants did not view all three robots.

First, while basic use of the Android phone was demonstrated, participants were not provided with any training on how to use the robot smartphone app or its capabilities. The popup message appeared shortly after the run began, as one of the first experiences with the app. Therefore, participants may have believed that any future information would also come in the from of popup messages.

Another related possibility is that some participants may have simply not been interested in viewing the information of some robots, as with two participants who only viewed two of the three robots (both did not view the "dead" robot).



(a) Learnability **(b)** Information **(c)** Control

Fig. 19. Robot smartphone app characteristics

The third possibility is that some participants may not have been able to figure out how to access the information about the other robots. Of the eight participants who only viewed one robot, five of them reported having no prior experience using Android devices.

People were able to use the app to identify solutions to problems and allocate their time more appropriately. Participants were much more likely to get two robots working when equipped with the app, with a combined time for robot cleaning that was over a minute and a half longer (on average) than without the app. During app runs, participants tended to focus most of their time and attention on the robot which needed to be reset first.

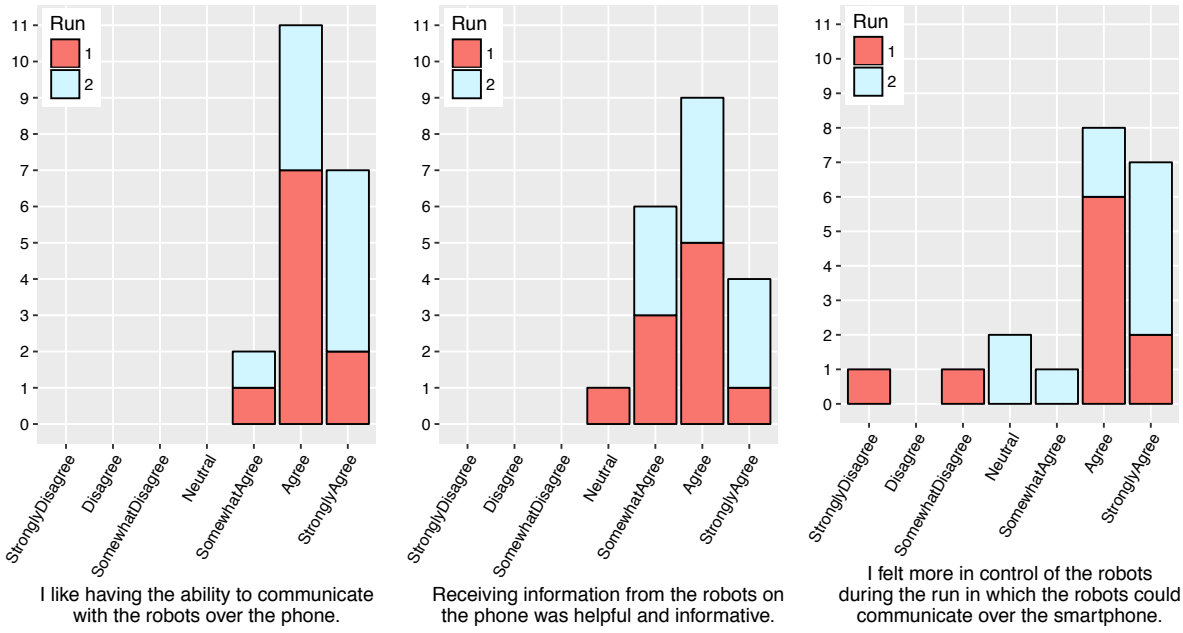
People using the app were also able to get the robots to clean for longer while also spending less time watching and physically interacting with them. We had predicted that this kind of behavior would lead to better performance since participants would have more time to score points in the balloons game and retain a higher percentage of their score. However, despite gaining an average of 30 additional seconds to play the balloons game when the app was enabled, people using the app actually spent significantly less time playing the balloons game, causing their overall performance to be about the same. Much of the lost time was spent using the app. One potential explanation for this behavior is that the app's ability to communicate with and control the robots had a strong novelty effect on participants, leading them to spend more time with it than with the balloons game. This explanation is supported by the fact that only 3 participants had previously used a robot vacuum cleaner.

6.3 H3 is Supported by the Results

Participants liked having access to the app, the additional information it provided, and its controls over the robots' built-in interfaces. All of the participants reported that they liked being able to communicate with the robots through the app, and all but one (who was neutral) thought the app was helpful. The majority of participants said they felt more in control using the app. The majority also felt much more confident about understanding what the robots were doing and what needed to be done to make the robots work. According to the NASA TLX questionnaires, people felt less discouraged while using the app. Unsurprisingly, people also reported higher levels of satisfaction with using the robots during the run with the app. According to workload data from the NASA TLX questionnaires, the app reduced people's mental and physical workloads.

6.4 Effects of Run Ordering

Some of our results show evidence of an ordering effect. For example, participants who used the app in the first run were able to apply knowledge from their first experience during their second run (e.g. resetting the robot



(a) Preferred app **(b)** App helpful **(c)** Feeling of control

Fig. 20. Robot smartphone app preferences

which needed help). Additionally, the 100% increase in participants who could reset the robot in the second run with the app who had previously been unable to without the app suggests that they had a better understanding of the information the app was providing due to their prior experience.

NASA TLX responses from after each run also show ordering effects. Some participants found the app to be distracting (4/20) or its messages confusing (7/20), with most of the complaints coming from people who used the app during the first run. Participants who used the app during their second run had the additional context on the task. The results from a similar question asked at the end of the experiment support this theory; in that question, all but one participant said they thought the information from the app was helpful.

7 Future Work

Perhaps the most important limitation of this study is that the participants were not representative of bystanders, who we consider to be an important target audience for this work, ultimately. Instead, participants were acting as operators or supervisors, with the robots' goals being aligned with their own goals. That said, given the lack of training provided, participants did have several characteristics of typical bystanders: a lack of familiarity or prior experience with the robot platform and a lack of training with its user interfaces. In order to measure how the feedback could assist people who are true bystanders to robots, we would test situations in which people must interact with a robot without previously having been informed that a robot would even be involved, but such experiments are difficult to construct. An alternative, less deception oriented experiment might ask people to find and help an autonomous robot carry out a task it has been assigned without prior knowledge of exactly where the robot is, what it looks like, what it is doing, or how to communicate with the robot. We believe that the results of our study are promising for these future bystander experiments, as we have shown that the app is able to convey status information and methods for assisting the robots.

Another limitation of this work is the lack of explicit comparisons of the effectiveness of *push* vs *pull* interactions. Both interaction styles were used during the experiment; however, they were directly paired with a single type of problem, and always occurred in the same order with participants receiving a *push* interaction popup message shortly after the beginning of the run. Testing the difference in effectiveness of these interaction styles is warranted. Finally, one of the most powerful applications of this work is its potential to be

relevant for communicating with a wide variety of different kinds of robotic platforms. Further testing with different kinds of robots, including drones and self-driving cars, is necessary to determine if this technology would be suitable as a ubiquitous method for communicating with publicly deployed autonomous robots.

8 Conclusions

The results of this work support the use of smartphones as a ubiquitous interaction method to allow untrained users to communicate with autonomous robot systems. Participants were able to use the app without training, and reported that it was easy to learn and use. All of the participants preferred having access to the app, and all but one said the app was was helpful. Participants were able to retrieve information about nearby robots, and could distinguish the source of the information despite similarities in the robots' appearances. With the app, participants felt more confident they understood what the robots were doing and were more satisfied with the robots' performance. Although participants' experiment scores did not improve with the use of the app, their behavior (specifically, spending less time watching and physically interacting with the robots, and getting more robots working for longer periods of time) created the potential for improved performance. While additional experimentation is needed to better understand the differences between the *push* and *pull* interaction methods and how bystanders might use the system, these results are a promising first step towards building communication between people and the increasing number of autonomous robots in our society.

9 Acknowledgements

The research in this paper was supported in part by the National Science Foundation under awards IIS-1552228 and IIS-1552256. Thanks to Chuta Sano and Christopher Munroe for their assistance with aspects of the experiment. Daniel Brooks was a doctoral candidate at the University of Massachusetts Lowell when this work was conducted.

References

[1] D. J. Brooks, A. Shultz, M. Desai, P. Kovac, and H. A. Yanco. Towards state summarization for autonomous robots. In *AAAI Fall Symposium Series*, 2010.

[2] J. Carlson and R. R. Murphy. How UGVs physically fail in the field. *IEEE Transactions on Robotics*, 21(3), 2005.

[3] E. Cha, M. Mataric, and T. Fong. Nonverbal signaling for non-humanoid robots during human-robot collaboration. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction*, pages 601–602, March 2016.

[4] M. Desai, P. Kaniarasu, M. Medvedev, A. Steinfeld, and H. Yanco. Impact of robot failures and feedback on real-time trust. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*, Tokyo, Japan, 2013.

[5] V. Groom, J. Chen, T. Johnson, F. A. Kara, and C. Nass. Critic, compatriot, or chump?: Responses to robot blame attribution. In *Proceedings of the 5th ACM/IEEE International Conference on Human-robot Interaction*, 2010.

[6] S. G. Hart and L. E. Staveland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, 52:139–183, 1988.

[7] Y. Hiroi and A. Ito. ASAHI: OK for failure: A robot for supporting daily life, equipped with a robot avatar. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*, pages 141–142, 2013.

[8] H. Hüttenrauch and K. Severinson Eklundh. To help or not to help a service robot: Bystander intervention as a resource in human–robot collaboration. *Interaction Studies*, 7(3):455–477, 2006.

[9] iRobot. *Features of the iRobot HOME App*, 2017 (accessed December 27, 2017).

[10] iRobot. *iRobot Robot Vacuums*, 2017 (accessed December 27, 2017).

[11] iRobot. *Resetting the iRobot HOME App "Care" status*, 2017 (accessed December 27, 2017).

[12] P. Kaniarasu and A. Steinfeld. Effects of blame on trust in human robot interaction. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, Aug 2014.

[13] T. Kim and P. Hinds. Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. In *Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication*, 2006.

[14] R. A. Knepper, S. Tellex, A. Li, N. Roy, and D. Rus. Recovering from failure by asking for help. *Autonomous Robots*, 39(3):347–362, 2015.

[15] M. K. Lee, S. Kiesler, J. Forlizzi, S. Srinivasa, and P. Rybski. Gracefully mitigating breakdowns in robotic services. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction*, March 2010.

[16] B. Lussier, R. Chatila, F. Ingrand, M.-O. Killijian, and D. Powell. On fault tolerance and robustness in autonomous systems. In *3rd IARP-IEEE and RAS-EURON Joint Workshop on Technical Challenges for Dependable Robots in Human Environments*, pages 351–358, Sept 2004.

[17] Neato Robotics. *Neato Botvac Connected*, 2015 (accessed December 27, 2017).

[18] Neato Robotics. *Neato Robot Vacuums*, 2017 (accessed December 27, 2017).

[19] S. Rosenthal, M. Veloso, and A. K. Dey. Is someone in this office available to help me? *Journal of Intelligent & Robotic Systems*, 66(1):205–221, 2012.

[20] G. Steinbauer. A survey about faults of robots used in RoboCup. In *RoboCup 2012: Robot Soccer World Cup XVI*, pages 344–355. Springer, 2013.

[21] D. Szafir, B. Mutlu, and T. Fong. Communicating directionality in flying robots. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 19–26. ACM, 2015.

[22] H. Yasuda and M. Matsumoto. Psychological impact on human when a robot makes mistakes. In *IEEE/SICE International Symposium on System Integration (SII)*, pages 335–339, Dec 2013.