Int J Soc Robot (2012) 4:117–129
DOI 10.1007/s12369-012-0138-y

Acquiring Accurate Human Responses to Robots' Questions

Stephanie Rosenthal · Manuela Veloso · Anind K. Dey

Accepted: 20 January 2012 / Published online: 18 February 2012
© Springer Science & Business Media BV 2012

Abstract In task-oriented robot domains, a human is often designated as a supervisor to monitor the robot and correct its inferences about its state during execution. However, supervision is expensive in terms of human effort. Instead, we are interested in robots asking non-supervisors in the environment for state inference help. The challenge with asking non-supervisors for help is that they may not always understand the robot's state or question and may respond inaccurately as a result. We identify four different types of state information that a robot can include to ground non-supervisors when it requests help—namely context around the robot, the inferred state prediction, prediction uncertainty, and feedback about the sensors used for the predicting the robot's state. We contribute two wizard-of-oz'd user studies to test which combination of this state information increases the accuracy of non-supervisors' responses. In the first study, we consider a block-construction task and use a toy robot to study questions regarding shape recognition. In the second study, we use our real mobile robot to study questions regarding localization. In both studies, we identify the same combination of information that increases the accuracy of responses the most. We validate that our combination results in more accurate responses than a combination that a set of HRI experts predicted would be best. Finally, we discuss the appropriateness of our found best combination of information to other task-driven robots.

Keywords Human-robot interaction · Asking for help · User studies

S. Rosenthal (✉) · M. Veloso
Computer Science Department, Carnegie Mellon University, Pittsburg, PA, USA
e-mail: srosenth@cs.cmu.edu

M. Veloso
e-mail: veloso@cs.cmu.edu

A.K. Dey
Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburg, PA, USA
e-mail: anind@cs.cmu.edu

1 Introduction

Robots use a variety of sensors to perceive and make inferences about their state and their environments. These sensors are often noisy, leading to uncertainty in the robot state which can result in errors while executing. As a result, human supervisors are often required to monitor robot performance and help reduce their uncertainty, even though full-time supervision is not scalable as we continue deploy more and more robots.

Our goal, instead, is for robots in our environments to identify when they are uncertain and proactively request assistance from the humans around them (*e.g.,* the robot's users [19, 24] or other humans available in the environment [26]). The proactive requests for help eliminate the need for expensive supervision. However, unlike prior approaches that assume humans are always knowledgeable about robots and state inferences when providing help, such as active learning [11, 22], learning by demonstration [3] and mixed-initiative and semi-autonomous robots [4, 5, 30], we cannot necessarily assume that non-supervisor humans in the environment will be knowledgeable about robots and always answer correctly [16]. This work focuses on how robots can increase the likelihood of receiving correct responses from non-supervisors.

In particular, we are interested in the types of robot information that help non-supervisors make more accurate inferences about the robot's state. We performed an extensive

human-robot interaction (HRI) and human-computer interaction (HCI) literature review to understand the types of information that other robots and devices have used to ask for help from humans. We found four common types of information that researchers in HRI and HCI have used (*e.g.,* [7, 14, 18, 28]):

- Context: The current sensor information related to the task (*e.g.,* features detected through vision or LIDAR);
- Prediction: The current inferred state (*e.g.,* the object detected with vision or the location of the robot),
- Uncertainty: The probability the inference is incorrect, and
- Feature feedback: The critical features from context used in inference (*e.g.,* the number of sides of the object or the carpet pattern near the location).

While different combinations of these four types of information are often used when requesting help from humans, little work has been performed to identify which combinations result in more accurate responses.

Our contribution is three-fold. After reviewing prior work on the types of information that is often provided when requesting help, we first contribute a novel two-part study design to evaluate which combination of the four types of information results in the most accurate non-supervisor responses. In the design, participants are asked to perform a task to limit their ability to supervise the robot. As our robot requests help on its own task, participants determine whether it is worth interrupting their task to respond. The robot is wizard-of-oz'd to ensure that the only difference between study conditions is the information provided; each participant receives the same questions at the same time during their task. After an initial test of many different combinations of information, our design includes a validation study to explicitly test our best found combination against a combination that HRI experts believed would result in the most accurate responses.

Our second contribution is the results of a shape recognition task and a localization task that use our study design. In the shape recognition task, a wizard-of-oz'd toy robot asks participants to identify the shapes of blocks they are manipulating. In the localization task, a remote-controlled real robot asks participants to identify their current location while giving a tour of the building. In both tasks, we vary the combinations of the four commonly used types of information that a robot provides to understand how such combinations affect the accuracy of participant responses. We find that providing all four kinds of information together—*context, prediction, uncertainty, and feature feedback*—most improved the accuracy of participants responses in both studies.

While this result may seem obvious, our group of HRI researchers predicted that a combination without a prediction and with only minimal context would result in more accurate responses. In a direct comparison with their selected combination, we validated that our best found combination shows a statistically significant improvement in accuracy. We thirdly contribute our combination as a guideline, validated in two domains and against HRI expert input, that can be used in new domains to increase non-supervisor accuracy.

2 Related Work

Human-human interactions are often grounded in the common references and experiences we have with others [10]. When we ask for help from other humans, these common experiences help clarify the question being asked and help us answer as accurately as possible. However, because humans may not share knowledge or references with robots, it has been suggested that robots should explicitly share their state information with humans as they act in the world [9]. We are interested in the types of information that could and should be shared with humans.

Researchers in human-computer interaction (HCI) and human-robot interaction (HRI) use different types of information when implementing requests for help on devices. We found that asking for help is common when there is uncertainty in inference (*e.g.,* recognizing or labeling objects in images [1] or localizing a robot [4]). While several different sets of guidelines have been proposed for what types of information devices should provide humans (*e.g.,* [7, 14, 18]), little work has studied which types actually increase response accuracy.

2.1 Task Domains

A variety of domains have used human help to reduce uncertainty when making inferences. In HCI domains, systems use user feedback to correct email and news article classifications [15, 23], handwriting recognition errors [29], and other context-aware inference errors [2]. Human computation and crowd-sourcing are also becoming increasingly popular ways to request labeled data from humans outside of any particular domain. For example, a game can use players' answers to label images for a search engine [1]. In robot domains, supervisors typically have an interface that shows robot and task state and allows them to control the robot's behavior (*e.g.,* [30, 34]). Robots have also autonomously requested localization help in offices [24], as well as help navigating both in offices [4] and outdoors [33].

In this work, we focus on two particular robot tasks that could require human help—shape recognition and localization.

Shape/Object Recognition Shape recognition (*i.e.,* identifying shapes of building blocks as a cube, cylinder, *etc.*)

Table 1 Operational definitions for each of the four types of information we focus on in this work, and examples of that information in our two task domains: Shape Recognition (*Shape*) and Localization (*Loc*)

State info.	Operational definition	Examples
Local context	The features around the area that the robot is trying to classify.	*Shape:* "You are working with the red and green blocks." *Loc.:* "I am near the kitchen."
Local+global context	The location of the local context in the state space.	*Shape:* "You are working with the red and green blocks on the top left" (of the tower). *Loc.:* "I am near the kitchen by the 7100 corridor."
Prediction	The most probable answer.	*Shape:* "Prediction is a rectangular prism." *Loc.:* "I think I'm at the red dot." (dot on map)
Uncertainty	Probability the inference is incorrect.	*Shape:* "Cannot determine the shape." *Loc.:* "Cannot determine the location."
Feature feedback	Ask the human for a set of contextual features that are indicative of the answer.	*Shape:* "What features describe the block?" *Loc.:* "Please describe the location."

is similar to camera-based object recognition that a robot might have to perform. In such recognition tasks, if a robot cannot determine the shape of an object that it is supposed to pick up, it may fail to complete its task. It could instead ask a human to identify the shape or object when it is uncertain (*i.e.,* asking "What shape is the red block?") in order to overcome the recognition failure and complete more tasks.

Localization Many mobile robots perform localization to determine where they are and how to navigate to a goal location. If a robot cannot determine its location, it may miss turns and have to backtrack down the hallway. We have shown that a robot can ask a human to identify its location on a map when it is uncertain (*i.e.,* "Can you point to where we are on this map?" (shows map)) to navigate more quickly and accurately to its goal [24].

2.2 Types of Information to Provide

We analyzed different systems that request help in different domains and categorized the types of information they provide to contextualize or ground those requests [10]. We found four popular categories of information also proposed by other researchers (*e.g.,* [7, 14, 18]):

- Context: The current sensor information related to the task (*e.g.,* features detected through cameras for vision or through LIDAR or WiFi for localization). We further divide this into local context sensor data around the inference and local+global context which additionally grounds the sensor data within the entire state space;
- Prediction: The current inferred state of the robot (*e.g.,* the shape of the object detected with vision or the (x, y) location of the robot);
- Uncertainty: The probability the inference is incorrect, and
- Feature feedback: The critical features from context used in inference (*e.g.,* the number of sides of the object or the carpet pattern near the location).

While these other researchers have also proposed including other information such as the current action that is being executed [7], acknowledgment of accountability to the humans in the environment [7, 14], and the costs and benefits of different user responses [18], we did not find these other types as prevalent in implemented systems.

Next, we provide operational definitions for each kind of information and provide examples of how to implement them in our two task domains (summarized in Table 1). We compare the accuracy of non-supervisor responses to robots' questions that include different combinations of the information.

Context Many robots and other applications provide humans with some contextual information about their sensor data before asking a question. However, some provide more contextual information than others. For example, a robot user interface to monitor speech recognition errors provides the audio that could not be recognized and a transcript of the conversational context [30]. Another robot provides no context at all about its current sensor readings when asking for help with localization and navigation in an office hallway (*e.g.,* [4]).

We define two kinds of contextual information: local context and local+global context. Local context are the features immediately around the state that the robot is trying to classify or infer. For example, in the speech recognition user interface described above, the local context is the audio recording of the sentence that is not recognized. The local+global context additionally contextualizes the local context in the entire state space (*e.g.,* the unrecognized sentence within the current conversation). In our shape recognition task, the local context is the color feature of the object in question and the local+global context is the location of the object in the image area (*e.g.,* top, left, bottom, right).

Prediction The prediction is the most likely state based on the inference. In speech recognition, the prediction is the

most likely sentence that was spoken [30]. An interface may automatically fill in fields in an online form or provide a prediction for which folder to sort an email into (e.g., [12, 15]). In our shape recognition task, the prediction is the most likely shape (i.e., cube or cylinder). Providing a prediction may reduce a user's work to respond because they only have to confirm an answer rather than generate it [13]. In this work, we test the accuracy of participant responses when the robot provides a correct prediction. Testing accurate predictions allows us to understand how people trust the robot and how much they are paying attention to what the robot says.

Uncertainty Many classification and inference algorithms give a measure of uncertainty—the probability that a prediction is inaccurate—in addition to the prediction itself. Studies of context-aware and recommender systems show that providing users with the level of uncertainty in predictions improves its overall usability (*e.g.,* [6, 21]), even if the system does not provide the exact uncertainty value [2]. For example, in the shape recognition task, the robot indicates that it is uncertain with the phrase "Cannot determine the shape."

Feature feedback We define feature feedback as requesting users list a set of contextual features that are most important for the inference. For example, in the shape recognition domain, feedback might include the number of edges or sides a shape has. It has been shown in both the active learning and HCI research that people are capable of providing useful feature feedback to a system. For example, in text classification domains, people were able to indicate not only the type of news article (sports, current events, *etc.*) but also keywords in the article that determine the type (*e.g.,* team or score for sports) [23]. People have also been able to successfully provide corrective feedback for handwriting recognition, email classification, and other domains (*e.g.,* [20, 27, 29]). We test whether asking people to provide this additional feedback influences the accuracy of non-supervisor responses to inference questions.

2.3 Combining the Information

Despite the common use of our four types of information, we found that different combinations of them have been used on different robots. For example, search and rescue robot interfaces for supervisors almost always include the robot's local context and inference predictions [34]. However, sensor uncertainty and feature feedback did not appear in interfaces, because supervisors implicitly also knew about the robot's uncertainty and were able to give feedback about the features without being asked.

In total, there are $3 \times 2 \times 2 \times 2 = 24$ different combinations of this information that could be provided. There are three ways to provide context: no context, local context, or local+global context. For each of those choices it could provide an inference prediction or not. For each of those six choices, it could provide uncertainty information or not. And finally, for each of those 12 choices, it could request feature feedback or not. In this work, we combine the types of information in the following order: (1) uncertainty, (2) context, (3) the question the robot wants answered, (4) prediction, (5) feature feedback. For example, when the robot in the shape recognition task asks about a block with all four kinds of information, it would say:

Robot: "Cannot determine the shape. You are working with the red and green blocks in the top left. What shape is the red block? Prediction is Rectangular Prism."
Human: Answers
Robot Follow Up: "What features describe this block?"

However, if the robot only asks with uncertainty and prediction (no context or feature feedback), it would say:

Robot: "Cannot determine the shape. What shape is the red block? Prediction is Rectangular Prism."
Human: Answers

In our studies, we explore the impact of these different combinations of information on the accuracy of non-supervisor responses. In the shape recognition task, we will test all 24 combinations to find the most accurate. In the localization task, due to robot and time constraints, we test only 5 combinations. While the exact statements are domain specific, they illustrate how we use the operational definitions and can be easily generalized to other similar applications. Next, we contribute our study design that we used to test the combinations of information.

3 Study Design

We define non-supervisors as humans who have a task to attend to and do not monitor the robot's progress. Because they are busy with their own task, they may not hear the information the robot might provide when asking for help, they may be rushed to answer, and as a result, their answers may be incorrect or they may not answer at all. However, despite the interruptions, some non-supervisors, such as the robot's users, have incentive to accurately answer questions in order for the robot to be able to complete its tasks for them. For example, visitors, who are escorted to meetings by a robot, may have incentive to answer questions about localization so that they can continue following the robot to their meetings [24]. Recent studies on email systems confirm that people are willing to be interrupted if there is a perceived benefit for them later [31, 32]. We are interested in combinations of information that improve response accuracy under these conditions.

We contribute a two-phase study design, namely an initial exploration phase to test many combinations of information and a validation phase to explicitly compare our best found combination with a baseline combination. In the initial exploration phase, we vary the combination of information that participants receive when the robot asks for help to understand how it affects the accuracy of their responses to the questions. Because there are too many combinations of information to test the statistical differences of each combination individually, we instead test the effect of each type of information averaged over all combinations. In our initial experiment, we show which types of information has a positive effect on the accuracy of the participants, either alone or in conjunction with other information in a between-subjects design. Using the between-subject design, we can understand how user accuracy varied through the experiment without confounding responses by presenting multiple different combinations in a short period of time. We measure the response rate and accuracy to the robot's questions.

Although we do not test the statistical significance of each combination, we perform a validation comparing the combination which includes all positive information from the initial phase against a predicted best combination from a set of HRI experts. The validation demonstrates that our combination is statistically more accurate.

The design is intended to mirror real-world conditions of asking non-supervisors including the incentives to answer while controlling for variations in the timing of the questions.

No Supervision In both the initial exploration and validation phases, participants are given a task to complete and limited time to complete it, preventing them from supervising the robot. They are told that they will only be judged on their task performance but that they can help the robot if they have time to complete the task.

Incentive to Answer Non-supervisors have incentive to answer questions despite the interruption, because they want the robot to perform tasks for them. In our study design, participants are told that the robot will interrupt their task to ask them for help if it is uncertain of predictions it is making. Helping the robot during their current task will improve their performance on a second task (which they are never actually given), but answering is optional as it may slow down their performance on their current task. During the study, participants must determine if they have time to answer the question without affecting their task performance. We measure the response rate to understand how participants evaluated the tradeoff.

Control of Question Timing The timing of a question may significantly affect response accuracy if the question is referring to something that the participant is doing at the time. In order to control the timing, the experimenter triggered the same questions during the same events in the task for all participants in all conditions (the robot was wizard-of-oz'd [17]). Controlling the timing ensures that the only difference between study conditions is the information the participants receive when being asked for help. The ground truth of what the robot is asking about is what the experimenter triggers the robot to ask about. We compare the participants' responses to the experimenter's ground truth to measure response accuracy.

Two Phase Validation The initial exploration phase is a between-subjects experiment. Participants are assigned to one condition of the study and receive a single combination of information for all questions asked. By comparing participants' responses in the different conditions, we can determine the best combination of information. After an initial study, our design includes a second within-subject validation study to explicitly test our best found combination of information against a combination that HRI experts predict will result in the most accurate responses. We chose the HRI expert combination to serve as our baseline instead of a baseline with no information, because we expect that a robot that asks for help would be implemented with their predicted combination. The validation serves to show that our results are an improvement over this baseline.

Next, we describe two experiments conducted using our study design.

4 Study Method

To investigate the impact of a robot providing different combinations of information when asking for help, we compared the accuracy of non-supervisor responses during both a shape recognition and a localization task. We first conducted the initial phase of the shape recognition task, testing all 24 combinations of information. At the same time, HRI researchers were brought together to come to a consensus on which combination of information they thought would result in the highest accuracy—which we call the HRI expert input combination. After finishing our 24-condition shape recognition task, we ran a shape recognition validation to directly compare our best combination to the HRI expert input. Finally, we conducted the localization task experiment to test our best combination from the shape recognition study against four other combinations. We show that our single best combination outperforms the other combinations in both domains.

4.1 Task Procedures

Questions and Information Combinations Before each study began, we generated the questions and information

(a) Robosapien V2 robot (b) A block structure (c) A block structure

Fig. 1 (**a**) The robot asked participants to indicate the shapes of blocks they were holding as they built the structures. (**b**) and (**c**) Examples of structures participants were asked to build out of multi-colored blocks

the robot provided based on the expected state at the time the questions would be asked. We, first, determined which sensors that would be used in the task (camera for shape recognition and a WiFi sensor readings for localization). Then, we chose the blocks and locations the robots would ask for help about and used our operational definitions to generate the information the robots would provide about them. We combined the information and questions in the following order: (1) uncertainty, (2) context, (3) the question the robot wants answered, (4) prediction, (5) feature feedback. For example, for the shape recognition task, when the robot provided all information it would say:

Robot: "Cannot determine the shape. You are working with the red and green blocks in the top left. What shape is the red block? Prediction is Rectangular Prism."
Human: Answers
Robot Follow Up: "What features describe this block?"

Table 1 outlines examples of each type of state information for each task, and Sect. 2.3 outlines the different combinations of information.

Shape Recognition Initial Task For the shape recognition task, we asked participants to build structures out of blocks while the robot tried to recognize the block shapes [25]. Our robot in this study, the RoboSapien V2 robot (Fig. 1(a)), contains a camera to track primary colors and LEDs that rotate towards the motion so that it appeared to be watching the participants build the structures. Upon arrival for the study, participants were randomly but evenly assigned to one of the 24 combinations of information, given an explanation of the study and signed a consent form. Before starting the task, participants were told that during their building task, the robot might ask them for help. The building task prevented the participants from supervising the robot. They could choose not to respond to the questions if they were too busy with building the structures, but they were told that answering questions would benefit them in a second related task (which we did not actually have them perform).

Participants were, then, given 50 colored blocks and four pictures of structures each containing 20–30 blocks to build in 12 minutes (Fig. 1(b) and 1(c)). When each of 8 predesignated blocks were picked up by the participant, the experimenter pressed a button to make the robot ask participants to identify the shape of a block they were holding in their hand. The participants were then given a chance to answer the questions verbally if they chose to. After completing the task, participants were given a survey about their experiences with the questions. Then, participants were told there was not enough time to conduct the second task and were dismissed after being paid.

Shape Recognition Validation While the initial phase was run, we sought advice from three members of the HRI community about which information they believe the robot should use when asking for help. The community members understood both the technical data that could be collected and the usability requirements necessary for effective communication to non-supervisors. We explained each type of information and how the information could be combined together. To achieve maximum accuracy, they suggested that the robot should provide *uncertainty, local context, no prediction, and feature feedback*, which we call the HRI expert input combination. They believed that longer sentences in the global context condition would make participants have to listen longer, interrupting them more. Additionally, they thought that participants would not believe the predictions if a robot was asking questions. We test the HRI expert input against our best found combination from the initial shape recognition task to validate that our best combination is better than what would commonly be implemented on robots that ask for help.

The shape recognition validation was conducted as a within-subject design with participants receiving questions both with our best combination of state information and with the HRI expert input combination. Participants were randomly but evenly assigned to the combination of information they would receive first. Subjects were given the same

(a) CoBot - front (b) CoBot - back

Fig. 2 CoBot from the front (**a**) and back (**b**). Participants walked behind CoBot so that they could see the messages and questions. CoBot spoke the questions through speakers below the laptop

shape recognition task instructions in the initial phase. When they finished building their first four structures in 12 minutes, they were given a questionnaire. Then, they were given a second set of four structures (the order of the groups of structures were randomly and evenly assigned between the two conditions) to complete in 12 minutes while the robot asked questions using the second combination of information. In total, participants performed two 12 minute tasks, filled out a survey after each task, and then filled out a final survey to compare the two question conditions. When they completed the third survey, they were paid and dismissed.

Localization Task Our robot CoBot (Fig. 2), a real custom-built mobile robot, is capable of autonomous localization and navigation and provides services such as tours to our computer science building. However, it can be uncertain of its location when using WiFi localization [8]. We have shown that if CoBot could ask for localization help from people in the environment as it navigates, it can avoid localization errors and speed navigation time [24].

In our localization task, participants were asked to walk around with CoBot while it gave a 15-minute tour of one floor of the building. These participants had never been on this floor of the building and thus could benefit from the tour. Upon arrival, they were randomly assigned to one of five conditions: (1) no information, (2) uncertainty and local+global context, (3) uncertainty and prediction, (4) uncertainty, local+global context, and prediction, and (5) uncertainty, local+global context, prediction, and feature feedback (our best found combination from the shape recognition task). In pretests, we found that local context was not enough information for people who had never seen our building before. Additionally, we include uncertainty in four conditions because it has previously [2] been found that users tend to trust agents more when they admit they are uncertain. We included all combinations of local+global context and prediction, because the context and predictions provide similar information in different ways. Our 5th condition tests our best combination which also includes feature feedback.

The experimenter remote-controlled the robot to each location in the building, triggering information about seven different laboratories, art installations, and views from the windows as it navigated. During the tour, the experimenter stopped the robot in 13 pre-defined locations to ask participants to indicate the robot's location on a map (Fig. 3). Participants were told that the robot would not be able to continue the tour if they did not help it. Because the experimenter was standing behind the participant while he/she was following the robot, the participants could not see the experimenter trigger the questions or control the robot and they believed the robot was moving autonomously. We used CoBot's uncertainty and predictions from its autonomous navigation to guide our decisions in where we triggered questions during the study. After the participant clicked on the map to indicate their location, the robot would continue navigating. After participants completed the 15-minute tour containing 7 places of interest and 13 questions, they were given a survey about their experiences with the robot. Upon completing the survey, participants were paid and dismissed.

4.2 Participants

Forty-eight Pittsburgh residents ages 18–61 (mean 27.6, s.d. 2.4) with a variety of occupations including students, bartenders, teachers, and salesmen performed the shape recognition task (37 subjects in the initial phase and 11 in the validation). Forty-two participants were included in the localization study all of whom were graduate or undergraduate students at Carnegie Mellon University who had not spent time in our new computer science building. Only a few participants (15%) had experience with machine learning technology, and all spoke fluent English.

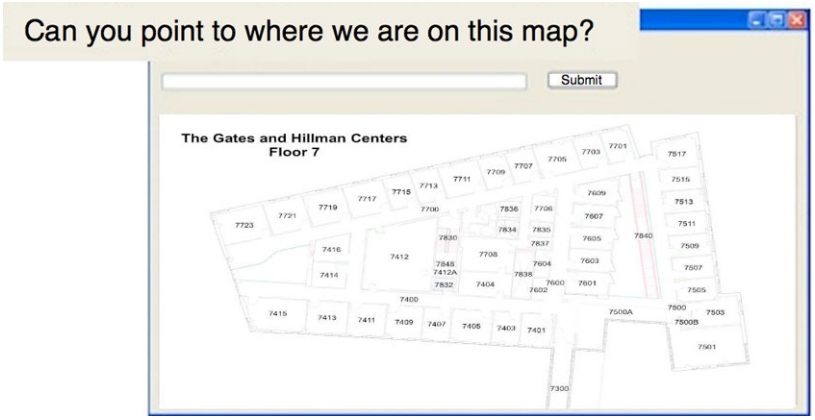

Fig. 3 CoBot stopped in 13 locations to ask participants to indicate their current location by clicking locations on its user interface

4.3 Measures

Because a robot agent would benefit more from correct answers to questions rather than incorrect ones, we assessed the non-supervisor responses to the questions primarily based on correctness. The responses in the shape recognition task were classified as a binary value: correct or incorrect. The responses in the localization task were measured as the Euclidean distance from the true robot location to the location the participants clicked on the map. We also gave surveys to all subjects about their opinions of the robots, asking questions including whether they found the applications to be annoying.

Shape Recognition Initial Task and Validation Participants' responses were classified as correct answers if their last answer (some users changed their minds) was correct and incorrect otherwise. For example, if a subject disagreed with the prediction, but gave an equally correct answer, it was classified as correct. Synonyms were determined to be correct as long as they were not too vague. For example, "rectangle" was considered a synonym to "rectangular prism" but "square" and "cylinder" were not.

Localization Task Participants clicked on a map displayed on the robot's screen to indicate their current location, and the (x, y) locations of their clicks were logged in order to determine the Euclidean distance to the actual robot location (Fig. 3). These mouse clicks could be used directly by the robot by translating the pixel coordinates into (x, y) coordinates in the building, making it an ideal way to ask for help. Each pixel is equal to about 4 inches and a hallway in the building is 15 pixels across. The mouse clicks were deliberate as participants often considered which pixel to press within a 1–2 pixel granularity. We recognize Euclidean distance does not distinguish incorrect hallways or inside offices as worse than a click in the appropriate hallways. However, our data indicates that when these errors occur, they are at large distances from the true location anyway.

Surveys After completing any of the tasks, participants were given questionnaires on their subjective experiences with each technology. They were asked about whether they thought the robot's questions were annoying and whether they found each dimension particularly useful. Responses were coded as either "Yes" or "No" because participants were not exposed to their combination of information many times and thus using a Likert scale to understand information preferences would not produce valid results. Participants were also asked whether it was easy or hard to answer the questions on a Likert scale from 1 (very easy) to 5 (very hard).

5 Results

We analyze the results of our studies to determine the combination of the four types of information that results in the most accurate responses. We find the same combination results in the highest accuracy in both domains. We will compare our shape recognition results with the HRI expert input to show that our combination improves accuracy a statistically significant amount.

5.1 Shape Recognition Initial Task

The robot asked all subjects at least 5 out of the 8 possible questions, due to some subjects running out of time. There was no significant difference in the number of questions answered for any particular combination of state information. Six percent of the questions that were asked were ignored due to the primary task. Seven participants skipped at least one question with two participants accounting for nearly half of the skipped questions. Of the answered questions, participants had an average error rate of 16.4% (s.d. 25%). This high standard deviation indicates that many (15) participants answered all questions correctly while several had very high error. We performed an ANOVA with the F statistic to test for ordering effects of whether the question number affected

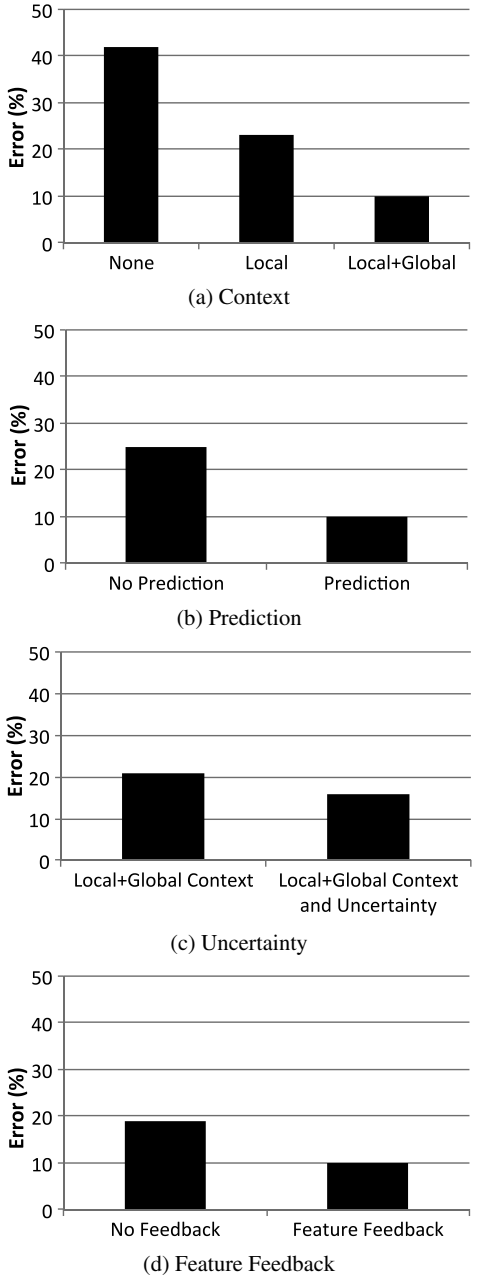

Fig. 4 (**a**) The more context the robot provides, the higher the accuracy of the participants' responses. (**b**) When the robot includes a prediction, the participants answer more accurately. (**c**) When the robot provides at least local context, the accuracy increases when the participant also receives uncertainty information. (**d**) When the robot asks for feature feedback, the participants answer more accurately

the participant accuracy. We found that there was no order effect and the accuracy did not change over the eight questions ($F(7, 170) = 1.70,\ p > 0.05$). McNemar tests with the χ^2 statistic were used to analyze the significance of the categorical response (correctness) against the categorical independent variables (our four types of information).

We analyzed the effects of each individual type of information on the proportion of correct answers the robot received. Figure 4(a), 4(b), 4(c), 4(d) show the percentage of questions that were incorrectly answered for context, predictions, uncertainty, and feature feedback, respectively. Subjects made statistically significantly fewer errors as they were given more context, dropping from 42% (none) to 23% (local) to 10% (local+global) ($\chi^2[2, 2] = 8.61,\ p < 0.02$). Subjects made significantly fewer errors when they received predictions (10%) compared to when they did not (25%) ($\chi^2[1, 1] = 3.59,\ p < 0.05$) and made fewer errors when asked about feature feedback (10%) compared to when they were not (19%) ($\chi^2[1, 1] = 4.05,\ p < 0.05$). There were no significant effects of uncertainty alone, but we found a significant paired effect of uncertainty and context reducing the error from 21% to 16% with local+global context and no significant difference in error without context ($\chi^2[2, 2] = 5.98,\ p < 0.05$). There were no other significant effects. Overall, we find that providing all four types of state information—local+global context, prediction, uncertainty, and feature feedback—increases the accuracy of non-supervisor responses. We will refer to this combination of information as our guideline for robots to ask non-supervisors for help and compare it to the HRI expert input next.

Subjects did not find any combination of dimensions more annoying than the others. Of the participants who received feature feedback, predictions, uncertainty or contextual information (local and local+global), 35%, 64%, 37% and 71%, respectively, found them to be useful.

5.2 Shape Recognition Validation

We compared the responses of participants in a within-subject design when the robots asked questions with the HRI expert input (local context, uncertainty, and feature feedback) to our guideline (local+global context, prediction, uncertainty, and feature feedback). T-tests were used to analyze the significance of the categorical response (correctness) against the two combinations of information (expert input and our guideline). There was no significant effect in the ordering of the conditions ($t[186] = 0.00,\ p > 0.05$). Figure 5 shows the percent of questions subjects answered incorrectly for each condition. There are significant effects of the combination on the proportion of correct answers subjects gave. Subjects provide significantly more correct answers (2.22% error) to the robot's questions when using our guideline compared to the expert input (15.63%) ($t[186] = 10.05,\ p < .01$).

Participants were asked whether they thought each kind of information was useful in helping them to answer the robot's questions. Subjects only scored the two systems differently for the contextual information dimension. While six participants gave our guideline combination (with local+global context) a score of 5 (very useful) for contextual information, only two participants gave the HRI expert

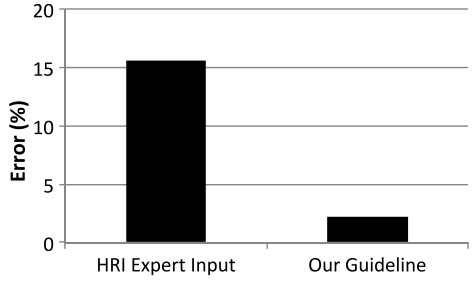

Fig. 5 Participants made significantly fewer errors when the robot provided our guideline combination compared to the combination determined by HRI expert input

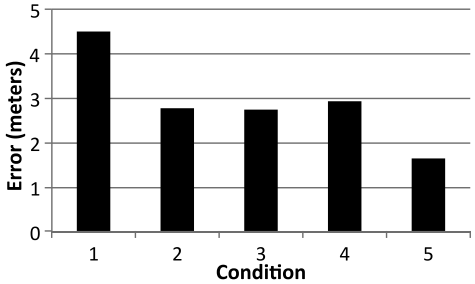

Fig. 6 The localization task had 5 conditions: (**1**) no state information, (**2**) uncertainty and local+global context, (**3**) uncertainty and prediction, (**4**) uncertainty, local+global context, and prediction, and (**5**) (our guideline) uncertainty, local+global context, prediction, and feature feedback. Participants who received our guideline combination of information responded with the least error

combination (with local context) the same score. However, a t-test shows no statistical difference between local context (3.46 average score) and local+global context (4.15) ($t[13] = 1.39,\ p > 0.05$). Participants rated our prediction on average 3.69 which is more positive than neutral, but we could not compare this to the expert condition which did not receive predictions. Subjects rated our uncertainty and the expert condition uncertainty (which were the same), 2.67 and 2.77 respectively ($t[13] = 0.18,\ p > 0.05$). Similarly, participants rated the feature feedback (which were the same) identically at 2.66 ($t[13] = 0.0,\ p > 0.05$).

Subjects were given another survey at the end of the experiment asking which system they preferred, which they thought was smarter, and which learned more (although the robot did not actually learn during the experiment). On all three survey questions, our guideline scored higher. Twelve out of fourteen respondents preferred our guideline over the expert input, eleven thought ours was smarter, and ten reported they thought ours learned more.

5.3 Localization Task

During the localization task, we collected the clicks on the map for each participant and calculated the Euclidean distance from the clicks to the actual robot location. Because the distribution of these distances was skewed, we performed a log transformation to normalize the data. We, then, analyzed the results of the localization test of log distances with a mixed model with participant ID as a random effect and the question condition as a fixed effect analyzed using the F statistic. Our results show there are statistically significant differences between the five conditions ($\mathrm{F}(4, 38.53) = 3.93,\ p < 0.001$). We used contrasts to analyze whether there were statistically significant differences between our guideline (condition 5) from the shape recognition study and the other four conditions tested. Running 4 contrasts means that statistical significance is determined at the level of $p < .05/4$.

Although we analyzed the log distances, we report the true distances in meters for clarity (Fig. 6). Participants who received no state information clicked further away from the robot's true location (4.5 meters) compared to those who received our guideline (1.65 meters) ($\mathrm{F}(1, 38.45) = 22.17,\ p < 0.001$). Participants who received only uncertainty and local+global context or uncertainty and predictions clicked 2.76 and 2.74 meters respectively from the true location, a marginally significant difference ($\mathrm{F}(1, 38.9) = 3.18,\ p = 0.082$) ($\mathrm{F}(1, 38.7) = 3.78,\ p = 0.059$). While our guideline shows a 1 meter improvement to these two conditions, there was a larger range of click distances for these conditions leading to only marginal significance. Finally, participants who received local+global context, uncertainty, and prediction clicked significantly further from the true location (2.94 meters) than those with our guideline ($\mathrm{F}(1, 37.6) = 8.17,\ p < 0.001$).

6 Discussion

Our results show that we were able to find a combination of information for robots to provide non-supervisors to increase the accuracy of their responses. Additionally, we were able to validate this combination of information in a second domain and against HRI expert input. Next, we discuss the impact of each kind of information on the human as well as the impact to the robot of providing this information.

6.1 Human Benefits of State Information

Interestingly, the combination of information that resulted in the most accurate responses for the non-supervisors is the same as the combination found useful for supervisors even though all the information was not found on supervisors' interfaces [34]. We found that participants in the shape recognition study rated the context and the prediction as useful to helping them respond, while supervisors similarly include the two types of information in their interfaces. Additionally, while neither our participants nor the supervisors rated the

uncertainty and feature feedback as not useful, both types of information were found to increase the accuracy of responses. Contrary to HRI expert intuition about which information increases the accuracy of responses, our result shows that all types of information have an impact on the non-supervisor. However, we still believe that a robot will not need to provide all information to supervisors, shortening the length of each required question, because they implicitly know the uncertainty and feature feedback information.

Context and Prediction The supervisors and non-supervisors use contextual information and prediction to focus their attention on what the robot is asking about. In the shape recognition task especially, where the participants were not shown the camera view of the robot, the full (local+global) description of where the robot was looking was useful to help participants find the block in question. Although the prediction was always correct, participants often did not trust it. The contextual information was used by the non-supervisors to check that the robot's prediction was consistent with the context it was providing. Additionally, the subjects' high rating of the predictions indicates that they listened to the predictions despite being busy with their primary task. A robot with less accurate predictions would need to focus more on providing contextual information to help people determine the accuracy of the prediction.

Uncertainty Although non-supervisors were frequently interrupted with questions in their structure building task and in the tour, they almost always answered the questions when it was prefaced with uncertainty information. Although these interruptions slowed them down, when the robot said explicitly that it was uncertain, the participants felt they should answer the question. We found a significant interaction of uncertainty and context in our analysis marked by improvement in accuracy with high levels of context, which confirms previous findings that users tend to trust or rely on systems more when the system displays uncertainty information [2]. However, when participants were asked whether they valued uncertainty, they did not remember if they had received the uncertainty information and did not report it as useful. We believe that participants underestimated how much they were using the uncertainty in their predictions.

Feature Feedback When participants were asked to provide feature feedback about their response, they sometimes changed their labels to the correct answer when they thought about why they chose the particular label. While it may be difficult for a system to incorporate such freeform feedback as we allowed, it has been shown that feature feedback can improve classifier accuracy [23]. Additionally, and perhaps more importantly, we have shown that the robot will benefit from increased response accuracy just by asking the question and irrespective of using the response.

6.2 Errors

We were initially surprised by the number of errors that participants made in both domains. However, upon further examination of the data, we found that there were two main causes of errors among our participants in each task. In the shape recognition task, the robot asked each question about the block the participants were currently holding in their hands. The participants often picked up multiple blocks at a time, causing mix-ups in shape when they were not paying attention. Additionally, participants continued building while the robot asked for help. If they put down their block and picked up another one of the same color, sometimes they would respond with the shape of the latter block although the robot started asking earlier. While these two problems could have been solved by requiring the participant to stop what they were doing to listen to the robot, it is unlikely that non-supervisors would stop what they were doing for the robot in real world situations. We believe, therefore, that our shape recognition results reflect real world scenarios at the cost of increased numbers of errors in many conditions. Our validation results show that using our guidelines, the error rate drops to 2% even when participants do continue working during the question.

In the localization study, the robot stopped moving to allow the participants to click on its attached laptop. Realistically, an error of 2.5 meters or more, as we received in all conditions except our best found combination, would not resolve the robot location uncertainty around an intersection to know whether to turn now or continue straight for another meter before turning. The two main causes of error in the localization task were due to (1) lack of knowledge of the building and (2) misunderstanding the robot's question. When participants did not know the building, they often found it hard to read the map even with every room labeled. Participants would often click on the correct corridor of rooms but did not focus their clicks close to specific room they were nearest to, resulting in clicks further down the hall away from the true robot location. Additionally, participants who found the room sometimes would click on the room itself rather than their location in the hallway. While this is an interesting response, it would result in large localization errors on a real robot.

Both of these errors were greatly reduced using our combination of information. The participants in our guideline condition had an average error of 1.65 meters, roughly the width of the hallway. While the robot can account for such error in its sensors, responses with larger errors would be difficult to use because the questions often occurred near hallway intersections when the robot is uncertain of whether to turn yet. The predictions on the map indicated to the participant to click in the hall instead of in a room. Most importantly, the feature feedback question resulted in participants looking around at room numbers more than in other

conditions. This heightened awareness likely impacted the responses the most in our guideline condition.

6.3 Computation Required to Calculate State Information

As we are motivated by task-driven robots that interact with people in the environment, we acknowledge that computation is limited on these robots and generating these questions may sometimes not be possible. We aimed to use information that was largely already calculated or known by the robot in order to reduce the computational requirements of asking for help. However, with limited resources, we found that a robot can increase the accuracy of its responses most with the least computation by providing at least local context, and also providing uncertainty and asking for feature feedback. We have found the most significant increases in accuracy when adding additional context, and suggest maintaining at least local context when asking for help. When CoBot provided context, participants' errors dropped from 4.5 meters to 3 meters to the robot's true location. Providing uncertainty information and asking for feature feedback both increase accuracy without having to generate any new information. The feature feedback, in particular, requires that the non-supervisor be more alert of the robot and the environment and results in more accurate responses, dropping localization error significantly from 3 meters to less than 2 meters.

7 Conclusion

When a robot asks for help, it must ground the human with information about its state in order to help them understand what the robot is inferring. While supervisors are grounded in the robot's current state because they are constantly monitoring it, people in the environment have no a priori knowledge about the robot. However, requiring supervision for each robot is expensive and not scalable so robots will need to ask people in the environment for help.

The contribution of this work is three-fold. First, we contributed a study design to test whether a robot's questions are understandable to non-supervisors. Second, we described the methods of two studies that use this design successfully to identify information that participants respond most accurately to. The same combination of information was found to increase the accuracy of non-supervisors the most in both studies. Third, we validated against a baseline combination of information from HRI expert input and contribute the combination as a guideline for the types of information a robot should provide when asking for help.

We believe our guideline can be used by other researchers given our validations. However, other types of information may also have an impact on how non-supervisors answer questions and need their own validation using our study design. Future work is needed to test our questions in long-term field studies among many more participants to understand the usability of robots that proactively ask for help as well as how answers change over time as people in the environment may get more familiar with a robot.

References

1. von Ahn L, Dabbish L (2004) Labeling images with a computer game. ACM conference on human factors in computing systems. In: CHI 2004, pp 319–326
2. Antifakos S, Schwaninger A, Schiele B (2004) Evaluating the effects of displaying uncertainty in context-aware applications. In: UbiComp 2004, pp 54–69
3. Argall B, Chernova S, Veloso M, Browning B (2009) A survey of robot learning from demonstration. Robot Auton Syst 57(5):469–483
4. Asoh H, Hayamizu S, Hara I, Motomura Y, Akaho S, Matsui T (1997) Socially embedded learning of the office-conversant mobile robot jijo-2. In: 15th international joint conference on artificial intelligence, pp 880–885
5. Asoh H, Motomura Y, Hara I, Akaho S, Hayamizu S, Matsui T (1996) Acquiring a probabilistic map with dialogue-based learning. In: Proceedings of ROBOLEARN-96, pp 11–18
6. Banbury S, Seldcon S, Endsley M, Gordon T, Tatlock K (1998) Being certain about uncertainty: How the representation of system reliability affects pilot decision making. In: Human factors and ergonomics society 42nd annual meeting
7. Bellotti V, Edwards K (2001) Intelligibility and accountability: human considerations in context-aware systems. Hum-Comput Interact 16(2):193–212
8. Biswas J, Veloso M (2010) Wifi localization and navigation for autonomous indoor mobile robots. In: ICRA 2010, pp 4379–4384
9. Clark H (2008) Talking as if. In: HRI'08: proceedings of the 3rd ACM/IEEE international conference on human robot interaction, pp 393–394
10. Clark H, Wilkes-Gibbs D (1986) Referring as a collaborative process. Cognition 22:1–39
11. Cohn D, Atlas L, Ladner R (1994) Improving generalization with active learning. Mach Learn 15(2):201–221
12. Culotta A, Kristjansson T, McCallum A, Viola P (2006) Corrective feedback and persistent learning for information extraction. Artif Intell 170(14–15):1101–1122
13. Eagle M, Leiter E (1964) Recall and recognition in intentional and incidental learning. J Exp Psychol 68:58–63
14. Erickson T, Kellogg WA (2001) Social translucence: an approach to designing systems that support social processes. ACM Trans Comput-Hum Interact 7(1):59–83
15. Faulring A, Myers B, Mohnkern K, Schmerl B, Steinfeld A, Zimmerman J, Smailagic A, Hansen J, Siewiorek D (2010) Agent-assisted task management that reduces email overload. In: IUI'10: proceeding of the 14th international conference on intelligent user interfaces, pp 61–70
16. Fong TW, Thorpe C, Baur C (2003) Robot, asker of questions. Robot Auton Syst 42(3–4):235–243
17. Green P, Wei-Haas L (1985) The rapid development of user interfaces: Experience with the wizard of oz method. Hum Factors Ergon Soc Annu Meet 29(5):470–474
18. Horvitz E (1999) Principles of mixed-initiative user interfaces. In: CHI'99: proceedings of the SIGCHI conference on Human factors in computing systems, pp 159–166

19. Lee MK, Kielser S, Forlizzi J, Srinivasa S, Rybski P (2010) Gracefully mitigating breakdowns in robotic services. In: HRI'10: 5th ACM/IEEE international conference on human robot interaction, pp 203–210
20. Mankoff J, Abowd G, Hudson S (2000) Oops: a toolkit supporting mediation techniques for resolving ambiguity in recognition-based interfaces. Comput Graph 24(6):819–834
21. Mcnee S, Lam SK, Guetzlaff C, Konstan JA, Riedl J (2003) Confidence displays and training in recommender systems. In: Proceedings of the 9th IFIP TC13 international conference on humancomputer interaction (INTERACT). IOS Press, Amsterdam, pp 176–183
22. Mitchell T (1997) Machine learning. McGraw Hill, New York
23. Raghavan H, Madani O, Jones R (2006) Active learning with feedback on features and instances. J Mach Learn Res 7:1655–1686
24. Rosenthal S, Biswas J, Veloso M (2010) An effective personal mobile robot agent through a symbiotic human-robot interaction. In: AAMAS'10: 9th international joint conference on autonomous agents and multiagent systems, pp 915–922
25. Rosenthal S, Dey AK, Veloso M (2009) How robots' questions affect the accuracy of the human responses. In: The international symposium on robot-human interactive communication, pp 1137–1142
26. Rosenthal S, Veloso M, Dey AK (2011) Is someone in this office available to help me? proactively seeking help from spatially-situated humans. Journal of Intelligent and Robotic Systems pp. 1–17
27. Scaffidi C (2009) Topes: Enabling end-user programmers to validate and reformat data. Carnegie Mellon Technical Report CMU-ISR-09-105
28. Shadbolt N, Burton AM (1989) The empirical study of knowledge elicitation techniques. SIGART Bull 108:15–18
29. Shilman M, Tan DS, Simard P (2006) Cuetip: a mixed-initiative interface for correcting handwriting errors. In: UIST'06: Proceedings of the 19th annual ACM symposium on user interface software and technology, pp 323–332
30. Shiomi M, Sakamoto D, Takayuki K, Ishi CT, Ishiguro H, Hagita N (2008) A semi-autonomous communication robot: a field trial at a train station. In: HRI'08: 3rd ACM/IEEE international conference on human robot interaction, pp 303–310
31. Stumpf S, Rajaram V, Li L, Burnett M, Dietterich T, Sullivan E, Drummond R, Herlocker J (2007) Toward harnessing user feedback for machine learning. In: IUI'07: proceedings of the 12th international conference on intelligent user interfaces, pp 82–91
32. Stumpf S, Sullivan E, Fitzhenry E, Oberst I, Wong W, Burnett M (2008) Integrating rich user feedback into intelligent user interfaces. In: IUI'08: proceedings of the 13th international conference on Intelligent user interfaces, pp 50–59
33. Weiss A, Igelsböck J, Tscheligi M, Bauer A, Kühnlenz K, Wollherr D, Buss M (2010) Robots asking for directions: the willingness of passers-by to support robots. In: HRI'10: 5th ACM/IEEE international conference on human robot interaction, pp 23–30
34. Yanco H, Drury JL, Scholtz J (2004) Beyond usability evaluation: analysis of human-robot interaction at a major robotics competition. Hum-Comput Interact 19(1):117–149

Stephanie Rosenthal is a Ph.D. student at Carnegie Mellon University in Computer Science. She conducts research in human-robot interaction, artificial intelligence, ubiquitous computing, and human-computer interaction. She received her B.S. in Computer Science and Human-Computer Interaction in 2007 and her M.S. in Computer Science in 2009 from Carnegie Mellon. Rosenthal is a National Science Foundation Graduate Research Fellow, a National Physical Science Consortium Fellow, and a Google Anita Borg Scholar. Rosenthal is a National Science Foundation Graduate Research Fellow, National Physical Science Consortium Fellow, Siebel Scholar, and Google Anita Borg Scholar.

Manuela Veloso is Herbert A. Simon Professor of Computer Science at Carnegie Mellon University. She researches in the area of Artificial Intelligence and Robotics. Veloso created and directs the CORAL research laboratory, for the study of autonomous agents that Collaborate, Observe, Reason, Act, and Learn, www.cs.cmu.edu/~coral. Veloso is IEEE Fellow, AAAS Fellow, and AAAI Fellow. She is the President-Elect of AAAI, the Association for the Advancement of Artificial Intelligence. She is also the recipient of the 2009 ACM/SIGART Autonomous Agents Research Award for her contributions to agents in uncertain and dynamic environments, including distributed robot localization and world modeling, strategy selection in multiagent systems in the presence of adversaries, and robot learning from demonstration. Veloso is the author of one book on "Planning by Analogical Reasoning" and editor of several other books. She is also an author in over 250 journal articles and conference papers. As of 2011, Veloso has successfully advised 23 Ph.D. students in Computer Science and Robotics.

Anind K. Dey is an Associate Professor in the Human-Computer Interaction Institute at Carnegie Mellon University. He holds a Ph.D. and M.S. in Computer Science, and a M.S. in Aerospace Engineering from Georgia Tech., and a Bachelor's Degree in Computer Engineering from Simon Fraser University. He conducts research on ubiquitous computing, mobile technologies, machine learning and human-computer interaction.