

How Robots’ Questions Affect the Accuracy of the Human Responses

Stephanie Rosenthal, Anind K. Dey, Manuela Veloso
Carnegie Mellon University
Pittsburgh PA 15213

{srosenth, anind, veloso}@cs.cmu.edu

Abstract— Asking questions is an inevitable part of collaborative interactions between humans and robots. However, robotics novices may have difficulty answering the robots’ questions if they do not understand what the robot is asking. We are particularly interested in whether robots can supplement their questions with information about their state in a manner that increases the accuracy of human responses. In this work, we design and carefully analyze a human-robot collaborative task experiment to measure humans’ responses and accuracies to different amounts of supplemental information. We vary the content of the questions along four dimensions of the robot state, namely uncertainty, context, predictions, and feature selection. Based on our results, we contribute guidelines on the effective combination of the four dimensions, under the assumption that the robot has no limitations on generating question context. Finally, we validate our guidelines against educated recommendations from the HRI community.

I. INTRODUCTION

Despite significant advances in robot navigation and obstacle detection, it is still difficult for robots to autonomously move about our buildings. Although we aim to contribute to the objective of autonomous robots, we realize that a mobile robot equipped with any number of sensors – even powerful ones – may not be able to process complex environments enough to act fully autonomously. Thus, our approach to robot autonomy is to support the ability of robots to proactively seek human assistance if needed.

When humans ask for assistance from other humans, it has been shown that the language used in a question can affect how people answer it, both in terms of the language used in the answer and the correctness of the answer [1]. Additionally, while humans share common experiences and can take each other’s perspective easily to answer questions, they also provide contextual information when necessary to supplement their questions [2]. Based on these phenomena, the social psychology and human-computer interaction (HCI) communities have developed guidelines on how to write survey questions and for techniques like focus groups, interviews, cognitive walkthrough, and pretests to help researchers iterate on and improve their questions, to reduce any possible ambiguities that people may find [3].

While there has been a large human-robot interaction (HRI) effort in providing techniques for building socially acceptable robots, and robots that understand human social cues (*e.g.*, [4], [5]), there has been little focus on the language robots should use when asking for assistance. In situations where robots are allowed to ask verbal questions of their human collaborators, there is often no information,

guidelines, or theory about how questions are formatted or asked (*e.g.*, [6], [7]). Instead, the robot’s questions are typically developed by the robot programmer and based on his or her intuition from human-human interaction of how questions should be asked. If the intuition is wrong and the robot’s questions are not clear enough, humans (especially non-experts in robotics) may not be able to understand the robot enough to give accurate responses.

The focus of our work is to understand how to make the robot understandable to robotics novices such that if it asks a human a question, the human can understand the robot’s perspective and then respond as accurately as possible. In this paper, we explore this concept in the context of a classification task, in which a robot attempts to classify blocks that it sees in its camera view by shape (*e.g.*, cube, cylinder), and we model the robot’s state for this task as a tuple of:

- **Context:** The current sensor information related to the task (*e.g.*, position, color, # of visible sides of the block);
- **Prediction:** The classification of the current context (*e.g.*, cube, cylinder);
- **Uncertainty:** The probability the classification is incorrect ($p \in [0, 1]$);
- **Features:** The critical features from context used to determine the classification (*e.g.*, # of sides).

When the robot has high uncertainty, it can ask a human questions about a shape while varying whether it explicitly includes information about each part of its state. We assume that the robot has “no limitations” in terms of its capabilities to generate questions. However, to eliminate all factors of uncertainty except for the content of the questions, we use a Wizard-of-Oz study with a robot embodiment and synthesized speech to ensure that the same questions are asked about the same blocks for each participant.

In this work, we present a method and subsequent guidelines for designing questions based on state information that robots can ask humans to maximize the accuracy of responses. First, we review how previous agents have explicitly explained their state to increase their usability. Then, we present our approach to studying subjects’ accuracy based on the robot’s questions and state information. Using the results from the study, we contribute a set of guidelines on the most effective combination of these four guidelines, for improving the accuracy of human responses to questions from robots. Finally, we validate these guidelines against current HRI

heuristics for robots asking questions and conclude that using our guidelines results in higher accuracy in human responses than using the heuristics.

II. RELATED WORK

Human-human interactions are typically *grounded* in common experiences that we all share or expect that we have shared [2]. When one of the participants in a conversation is not sure that the other has the relevant background to understand a topic, he explicitly states the context before explaining the topic. Because robotics novices do not share common bonds with robots, the intuition is that we should require robots to explicitly explain their state to humans to be understood [8]. We define the robot’s state in a classification task as the tuple (*context*, *prediction*, *uncertainty*, *features*). The robot represents its current sensor readings and any higher level inferences through *context*. The robot makes a *prediction* about how to classify the context, using a subset (either learned or static) of *features* \in context and maintains a probability *uncertainty* that it believes the prediction is incorrect. The robot can then use this state information to explain its current situation and request assistance, varying which dimensions it provides to the human.

Individually, each dimension of state has been shown to improve usability and even the accuracy of the responses of systems that require human interaction. Studies of context-aware, expert, and recommender systems all show that providing users with the level of *uncertainty* in a system’s predictions improves its overall usability (*e.g.*, [9]). Users are more willing to answer if they know it is asking *because* it is uncertain, rather than at a random time. Additionally, some results show that the exact uncertainty information is not necessary to improve performance, only the explicit mention that the system is uncertain [10]. Studies have shown that when the human and robot share a common frame of reference or *context* in the environment, they can communicate more effectively (*e.g.*, [11], [12], [13]). When a semi-autonomous robot asks a human for help, it requires additional effort on the part of the human to understand what is being asked, limiting their productivity (*e.g.*, [14]). In both HCI and HRI, there has been an effort for agents to proactively provide *predictions* of the next action to take and warnings that errors are about to occur in order to reduce the cognitive load of humans who are assisting them (*e.g.*, [14], [15], [16]). Finally, users can provide useful feedback to these systems about the *features* they would use to make a classification [17].

In situations where robots ask questions, there is often little support for how the content of the questions is generated. Some questions offer no state information at all (*e.g.*, [7]), while others provide all state information leaving the human to sort through it (*e.g.*, [14]). Other robots provide predictions but without uncertainty information or other context (*e.g.*, [6]). It is unclear which pieces of information are most important for the robot to give and which are excessive and in which situations. Additionally, because existing work does not vary the state information that users are provided, it is

unclear if there is other state information that would increase the accuracy of the users’ responses. The goal of our work is to present a method and subsequent guidelines for designing grounded questions based on state information that robots can ask human collaborators to maximize the likelihood of a correct response.

III. METHOD

We conducted a study to test the correctness of users’ answers to a robot’s questions based on the content of the questions asked. The robot was given a block shape recognition task, simulating camera-based object detection, in which it watched study participants build structures out of blocks. As the participants, who are typically good at identifying shapes, built their structures, the robot interrupted to ask for assistance if it did not recognize a block shape being used. We use an object recognition task because it is easy to assess ground truth and measure correctness: *i.e.*, confirming that the block the robot asks about matches the block the human describes and that the human describes that block correctly. To understand how a robot can ask questions to optimize the correctness of the responses it receives, we vary the content of the robot’s questions based on the dimensions presented above and measure both the correctness of the responses and the user’s opinions of the questions. While the content of our questions reflect the shape recognition task, the state information that we include and resulting guidelines can be easily generalized to other classification problems.

A. Task

The participants’ primary task is building structures out of blocks in a timed setting. They sit in front of the robot and are given a set of 50 wooden blocks, containing 6 different block shapes in 5 different colors. The subjects are given 4 pictures of block structures (Fig 1(a)), each composed of 20-35 blocks, to build in 12 minutes. The subject is told the robot is learning to recognize block shapes. Humans are better at recognizing objects and robots can leverage this ability by interrupting the human to ask questions about those objects. If participants have time while performing their

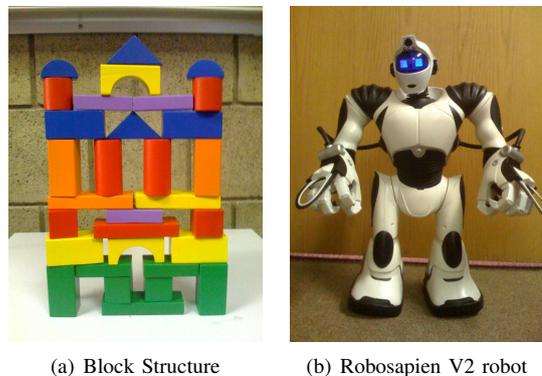


Fig. 1. The Robosapien V2 robot watches participants build structures and asks questions about the blocks it cannot recognize.

TABLE I
CONTENT DIMENSIONS, DESCRIPTION, AND EXAMPLE SENTENCES FROM THIS STUDY.

State Dimension	Description	Example Sentence
Context	The amount of contextual information the robot gives the human about what it is perceiving (none, local, global)	local (color): “You are working with the green and red blocks.” global (color and position): “You are working on the top left column of the structure with green and red blocks.”
Prediction	The robot tells the human which shape it predicts	“It might be a cube.”
Uncertainty	The robot notifies the human that it is uncertain of the block shape before asking the question	“Cannot determine the shape.”
Feature Selection	The robot asks the human why it is a certain shape (expected answer is a description of the block)	“Why?”

primary task, they can help teach a robot to recognize block shapes by answering its questions.

Participants are then reminded that answering the robot’s questions will slow them down in completing their primary building task. However, they are given a delayed incentive to answer the questions - they are told that they will be given a second building task that the robot will help them finish more quickly if they help it learn block shapes during the first building task. This tradeoff between future incentive and present delays may cause people to ignore the robot or rush to answer its questions instead of spending extra time to help it. However, although answering questions takes more of the human’s time, recent work in classifying email has demonstrated that a human may be more willing to give more time to provide the feedback, if it may decrease the possibility of additional interruptions [17] [18]. We model this tradeoff to get a more accurate understanding of how people answer questions under pressure of their own goals and motivation by future incentives. Participants are not actually given this second task.

A Wizard-of-Oz’ed RoboSapien V2 (Fig 1(b)) robot watches subjects build with blocks. The robot was preprogrammed to follow faces and red, green, and blue colored objects with a built-in color camera. Colored LEDs on the robot rotate towards the colored objects as they move across the robot’s field of vision. This behavior helps create the illusion that the robot is watching the participants build the structures. As subjects build the structures during the task, the robot interrupts with questions claiming it cannot determine the shape of the block being handled. The robot asks questions about 2 blocks for each structure the participants build, for a total of 8 questions. The questions were generated ahead of time based on specific blocks in the structures and the audio files based on these questions were played via speakers hidden behind the robot. This gave the appearance that the robot was asking the questions. Our Wizard-of-Oz design ensured that all participants were asked questions about the same blocks with the same timing.

After completing the task or after time has expired, participants are given a survey about their opinions and experiences with the questions the robot asked. After completing the survey, they are told there is not enough time to conduct the second building task with the robot’s help. Participants are then paid and dismissed.

B. Robot State and Questions

Although the robot is Wizard-of-Oz’ed, we model the robot’s state using the same four dimensions described in Section II. We assume that the robot has the capability to find a block in an image, determine context like its color, location, the number of visible sides and edges, *etc.*, and then make a prediction about the block shape with some uncertainty using a subset of features from the context. We then vary which dimensions the robot uses in questions for a 3x2x2x2 design. We study every combination of state dimensions the robot can provide while asking a question in order to determine which combination results in the most accurate responses from humans. Table I outlines each dimension, the information the robot provides if the dimension is included, and an example of how the dimension translates into the robot’s question for this task.

1) *Context*: We vary the amount of context the robot gave about the shape it asks about, to understand how much information is “excessive” for the subjects to receive. Participants receive one of three conditions: *no* contextual information, *local* context of the colors of nearby blocks, and *global* information including color plus the position of the block in the structure. Participants who receive global context heard statements like “*You are working on the top left part of the structure with green and red blocks.*” While the local context condition always provides enough information to find the block in question, the global information provides participants with an extra confirmation that they found the block the robot is referencing.

2) *Prediction*: Half of the users received a correct prediction from the robot about the block they are using (*e.g.*, “*It might be a cube.*”) and half do not receive any prediction. Because the robot always gives a correct prediction, we can measure how often a human trusts the robot prediction but cannot measure the impact of incorrect predictions.

3) *Uncertainty*: The robot tells half of the participants that it “*Cannot determine the shape*” and does not mention uncertainty to the other half. It is not necessary to provide exact uncertainty data, only to admit that it is uncertain [10].

4) *Feature Selection*: We vary whether the robot asks the human an additional question about which features it used to classify the block (*e.g.*, “*Why is the block a cylinder?*”). We expect a participant might say that the ends are circles, identifying the critical features they use to classify the cylinder. Although the robot in the future may be able to

parse the response, we do not and only study whether asking this question (and receiving the extra answer) affects the classification accuracy.

5) *Putting it Together*: When the robot is able to provide all of its state information including global context, it might ask a human the following question about a block:

Robot: “Cannot determine the shape. You are working on the top left part of the structure with green and red blocks. What shape is the red block? It is likely a cube.”

Human: Answers e.g., “It is a cube”

Robot Follow Up: “Why?”

Human: Answers e.g., “It has six sides - all squares.”

However, if the robot provides local instead of global context, predictions, and uncertainty, but does not ask about feature selection, it would say:

Robot: “Cannot determine the shape. You are working with green and red blocks. What shape is the red block?”

Human: Answers e.g., “It is a cube”; no robot follow up

Participants consisted of 37 Pittsburgh residents ages 18-61 with a variety of occupations including students, bartenders, teachers, and salesmen. Participants had no experience with robots and only a few (15%) had experience with technology that learns. All spoke fluent English. Each participant was assigned to one of the 24 (3x2x2x2) combinations of state information randomly but evenly. All of the conditions were performed by one person first and the conditions that were repeated were chosen randomly.

C. Measures

Because a robot would benefit more from correct answers than from incorrect ones, we assess the user responses to the questions primarily based on correctness. Users’ responses are classified as correct answers if their last answer (some users changed their minds) is correct and incorrect otherwise (Table II). For example, if a subject disagrees with the suggestion, but gives an equally correct reference, it is classified as correct:

Robot: It is likely a cylinder. **Subject**: No, it is a log.

As a secondary measure, we gave subjects surveys about their opinions of the questions the robot asked. This included whether they thought the robot’s questions were annoying and whether they found each of the four dimensions useful or not. Answers were coded as either “Yes” or “No” to each of the five questions.

TABLE II
CODING FOR CORRECTNESS

Category	Description	Examples
Correct	Agree	“Yes” or “That’s right”
	Correct label	“Cylinder” or “It is a log”
	Correct disagree	“No its a log”
Incorrect	Disagree	“No” or “No, its a *”
	Incorrect label	“cube” instead of “cylinder”
	No response	“I don’t know”

IV. RESULTS

Here, we present the results of the study we conducted to optimize the frequency of correct responses to the robot. We analyzed the data to find the overall best combination of dimensions, assuming the robot has unlimited capabilities. In the next section, we validate our findings against a combination of the dimensions that a group of HRI specialists chose.

A. General Statistics

The robot asked all subjects at least 5 of the 8 possible questions, with some subjects running out of time and others choosing not to answer. There was no significant difference in the number of questions answered for any dimension. 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 dimensions). We will use the significant results to develop guidelines about how robots with unlimited capabilities should ask questions and test these guidelines against input robotics experts gave about question content.

B. Quantitative Results

We analyzed the effects of each individual dimension on the proportion of correct answers the robot received. Figures 2(a) and 2(b) show the percentage of questions that were incorrectly answered for context and suggestions, respectively. Subjects made statistically significantly fewer errors as they were given more context, dropping from 42% (none) to 23% (local) to 10% (global) ($\chi^2[2, 2] = 8.61, p < 0.02$). Subjects made significantly fewer errors when they received suggestions (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 selection (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 ($\chi^2[2, 2] = 5.98, p < 0.05$). There were no other significant effects.

C. Qualitative results

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

D. Discussion

Our results show that all four dimensions show statistically significant differences in how people respond to the questions. Additionally, a majority of users found that the robot’s contextual information and suggestions were beneficial and helped them respond. Using these results, we can conclude that for a robot with *unlimited resources*, it is optimal to give humans information about **uncertainty, global context, prediction, and feature selection**. We add the provision that

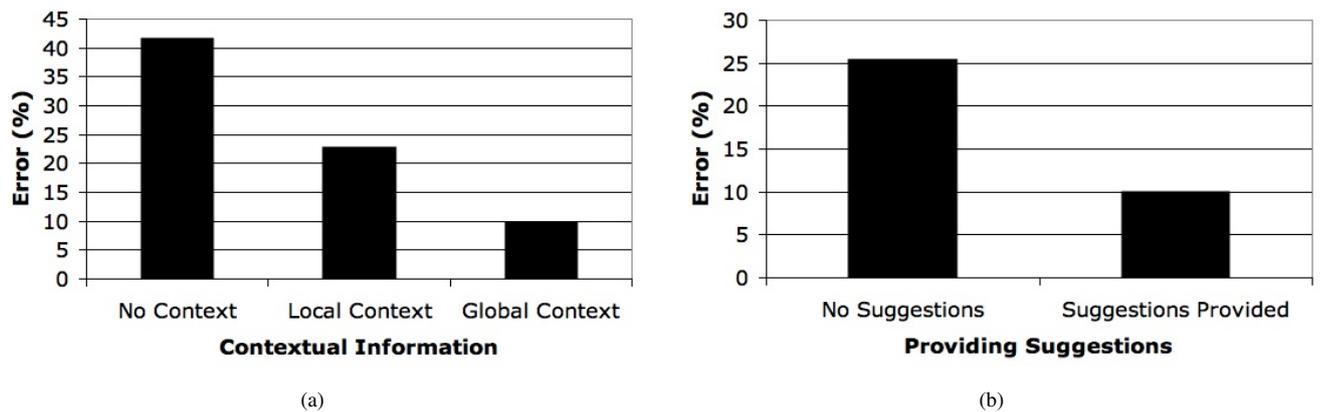


Fig. 2. Subjects made significantly fewer errors as they received 2(a) more contextual information and 2(b) suggestions.

they should be used only when necessary so that subjects are minimally annoyed.

V. VALIDATION

We asked for advice from 3 members of the HRI community about how robots should formulate their questions along our dimensions. Because robots' questions today are largely based on a researcher's input, their response acts as a baseline to test our guidelines against. We validate that our guidelines are at least as good as, if not better than, the heuristics used by these HRI community members.

A. HRI Heuristics

We explained each dimension to the 3 HRI researchers and explained how to combine the dimensions. The group discussed the dimensions and then reported their consensus. They concluded that along our dimensions, they would formulate questions that include **uncertainty, local context, no prediction, and feature selection**. This varies from our guidelines in two ways - the amount of contextual information and offering predictions.

B. Validation Method

We conducted a within-subject study in order to validate that our guideline improves the proportion of correct answers that people give the robot compared to the community input. Fourteen participants were told that they would be testing two different ways the robot learns from asking them questions. Similar to the first study, participants were told that they would complete both "learning" tasks to teach both robot programs and then later they would complete performance tasks to test how well the robot learned from them. As with the previous study, the robot was Wizard-of-Oz'ed and the robot was not actually learning. The order of the two conditions was randomized in addition to the order of the two sets of four block structures to be built. Each "learning" task was 12 minutes long and the subjects were given surveys after each task. Then they were told they did not have time to complete the performance tasks. The participants' answers were scored for correctness as before.

C. Validation Results

T-tests were used to analyze the significance of the categorical response (correctness) against the two sets of guidelines (ours and the community advice). There was no significant effect in the ordering of the conditions ($t[186, 1] = 0.00$, $p > 0.05$). Figure 3 shows the percent of questions subjects answered incorrectly for each condition. There are significant effects of the choice of guidelines 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 guidelines compared to the community advice (15.63%) ($t[186, 1] = 10.05$, $p < .01$).

Participants were asked, for each system, whether they thought each dimension 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 questions (with global context) a score of 5 (very useful) for contextual information, only two participants gave the community input questions (with local context) the same score. 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. On all three survey questions, our guidelines scored higher. Twelve out of fourteen respondents preferred the system using our guidelines over the system that used the community advice, eleven thought ours was smarter, and ten reported they thought ours learned more.

D. Discussion

Our results show that although participants gave a majority of correct answers to both robot systems, they provided significantly more correct answers for our guidelines. Using our systematic approach to testing the robot's questions, we found that robots that ask questions should also provide all of its state information along our four dimensions (uncertainty, global context, prediction, and features). Our subjects were able to take the robot's perspective easier and answer the robot's questions more accurately when they received all of the state information compared to when some of the information was left out. Despite our intuition that providing

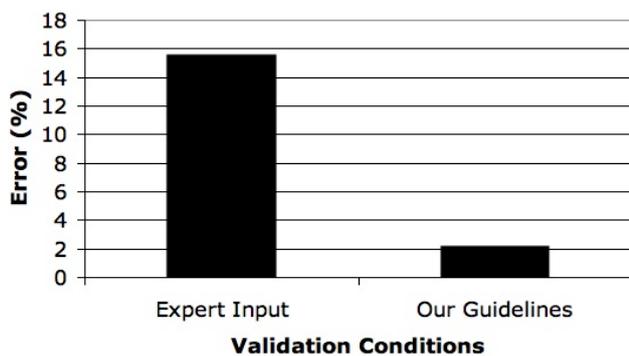


Fig. 3. Subjects made significantly fewer errors when they received our guidelines compared to the HRI heuristics.

more context (global) would be overwhelming or annoying to humans, our subjects found the additional confirmation that they had found the correct block more useful and less annoying than we expected. Additionally, although we found that our subjects did not always trust the prediction the robot gave, it did significantly increase the accuracy of subjects' responses.

Researchers can leverage these guidelines directly for use in their research without relying on their own intuition to generate questions for their robots to ask. We do, however, acknowledge that our state information may not be directly applicable to all robots or all tasks. Our guidelines assume that the robot has unlimited resources to calculate predictions, uncertainty, and the relevant features as well as unlimited resources to generate questions. We offer our method as an approach that other researchers can use to test their own questions for their own tasks to maximize the accuracy of the responses that their robots receive.

VI. CONCLUSION

Clear communication between robots and humans is necessary for collaboration. Humans must understand what their robot counterparts are asking about in order to give them correct information. Previous work in robot question-asking has not focused on validating the robot's questions to ensure that humans understand the questions and can correctly respond. The contribution of this work is two-fold. First, we formulate an approach to testing the validity of questions that a robot may ask. We show that this approach successfully identifies questions that humans respond well to, and provide the method for others to study the questions their robot asks. Second, we describe a set of guidelines derived from the results of our study for designing questions for robots to ask humans, validate them against current HRI heuristics, and show that our guidelines significantly increase the number of correct answers for classification questions that a robot asks.

This work focuses on a specific set of dimensions for classification problems. Additional work is needed to provide guidelines and validation of other types of questions a robot may ask and other dimensions that may affect how humans

understand and answer questions. Our work also does not focus on how robots can use the answers they receive from humans to incorporate them back into their models. Future work is necessary to investigate, for example, how to label the robot's real time sensor data using human feedback (especially verbal feedback).

REFERENCES

- [1] S. Presser, "Methods for testing and evaluating survey questions," vol. 68, no. 1, Mar 2004, pp. 109–130.
- [2] H. Clark and D. Wilkes-Gibbs, "Referring as a collaborative process," *Cognition*, Jan 1986.
- [3] R. Jeffries, J. R. Miller, C. Wharton, and K. Uyeda, "User interface evaluation in the real world: a comparison of four techniques," in *CHI '91: Proceedings of the SIGCHI conference on Human factors in computing systems*. New York, NY, USA: ACM, 1991, pp. 119–124.
- [4] A. Brooks and C. Breazeal, "Working with robots and objects: revisiting deictic reference for achieving spatial common ground," Jan 2006, pp. 297–304.
- [5] M. E. Foster, T. By, M. Rickert, and A. Knoll, "Human-robot dialogue for joint construction tasks," in *ICMI '06: Proceedings of the 8th international conference on Multimodal interfaces*. New York, NY, USA: ACM, 2006, pp. 68–71.
- [6] T. W. Fong, "Collaborative control: A robot-centric model for vehicle teleoperation," Ph.D. dissertation, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, November 2001.
- [7] H. Asoh, S. Hayamizu, I. Hara, Y. Motomura, S. Akaho, and T. Matsui, "Socially embedded learning of the office-conversant mobile robot jijo-2," in *Proceedings of 15th International Joint Conference on Artificial Intelligence (IJCAI-97)*, 1997, pp. 880–885.
- [8] H. H. Clark, "Talking as if," in *HRI '08: Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*. New York, NY, USA: ACM, 2008, pp. 393–394.
- [9] S. Banbury, S. Selcon, M. Endsley, T. Gorton, and K. Tatlock, "Being certain about uncertainty: How the representation of system reliability affects pilot decision making," 1998, pp. 36–39(4).
- [10] S. Antifakos, A. Schwaninger, and B. Schiele, "Evaluating the effects of displaying uncertainty in context-aware applications," vol. 3205. Springer Berlin / Heidelberg, 2004, pp. 54–69.
- [11] M. C. Torrance, "Natural communication with robots," Ph.D. dissertation, MIT, 1994.
- [12] E. Topp, H. Huttenrauch, H. Christensen, and K. Eklundh, "Bringing together human and robotic environment representations—a pilot study," Jan 2006, pp. 4946–4952.
- [13] L. Steels, "Evolving grounded communication for robots," *Trends in Cognitive Sciences*, vol. 7, no. 7, pp. 308–312, 7 2003.
- [14] M. Shiomi, D. Sakamoto, K. Takayuki, C. T. Ishi, H. Ishiguro, and N. Hagita, "A semi-autonomous communication robot: a field trial at a train station," in *HRI '08: Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*. New York, NY, USA: ACM, 2008, pp. 303–310.
- [15] S. Stumpf, X. Bao, A. Dragunov, and T. Dietterich, "Predicting user tasks: I know what you're doing," Jan 2005.
- [16] A. Faulring, K. Mohnkern, A. Steinfeld, and B. Myers, "Successful user interfaces for radar," *CHI 2008 Workshop on Usable Artificial Intelligence*, Jan 2008.
- [17] S. Stumpf, V. Rajaram, L. Li, M. Burnett, and T. Dietterich, "Toward harnessing user feedback for machine learning," Jan 2007, pp. 82–91.
- [18] S. Stumpf, E. Sullivan, E. Fitzhenry, I. Oberst, W.-K. Wong, and M. Burnett, "Integrating rich user feedback into intelligent user interfaces," in *IUI '08: Proceedings of the 13th international conference on Intelligent user interfaces*. New York, NY, USA: ACM, 2008, pp. 50–59.