Who Should I Blame? Effects of Autonomy and Transparency on Attributions in Human-Robot Interaction

Taemie Kim
taemie@mit.edu
The Media Laboratory
Massachusetts Institute of Technology
20 Ames Street, Cambridge, MA 02139, USA

Pamela Hinds
phinds@stanford.edu
Management Science & Engineering
Stanford University
Stanford, CA 94305, USA

***Abstract*— As autonomous robots collaborate with people on tasks, the questions "who deserves credit?" and "who is to blame?" are no longer simple. Based on insights from an observational study of a delivery robot in a hospital, this paper deals with how robotic autonomy and transparency affect the attribution of credit and blame. In the study, we conducted a 2x2 experiment to test the effects of autonomy and transparency on attributions. We found that when a robot is more autonomous, people attribute more credit and blame to the robot and less toward themselves and other participants. When the robot explains its behavior (e.g. is transparent), people blame other participants (but not the robot) less. Finally, transparency has a greater effect in decreasing the attribution of blame when the robot is more autonomous.**

I. Introduction

Robots are becoming increasingly common in our workplaces. The worldwide investment in industrial robots, for example, increased 18% in 2003 [1]. For many years, robots have been helping people by doing simple repetitive jobs. However, today's technology allows robots to take over more important and sophisticated jobs, some of which involve robots acting more autonomously.

As these sophisticated robots support humans in accomplishing their tasks, humans and robots may be collaborating more closely. This collaboration raises interesting questions: If they work together, who is responsible for the job? Who is to blame if something goes wrong?

These questions were raised as a result of an ethnographic study of an autonomous, mobile delivery robot deployed at "Community Hospital" located in an agricultural area of Northern California. This was a 20-month study done in 2002~2003 by a group of researchers including one of the authors. The researchers observed interactions between the robot and workers in the hospital and conducted interviews with some of the workers. The robot was a Pyxis HelpMate and its main function was to deliver medication from the pharmacy to nursing units around the hospital. The robot was able to navigate through hallways, ask for specific medications and call the elevator on its own.

From our analysis, an interesting pattern of interaction emerged. When an unexpected situation occurred, people were easily confused and did not know who was to blame: the robot, themselves or the other workers in the hospital who interacted with the robot. In some cases, nurses would attribute incorrect blame or too much responsibility to the robot or other nurses. In the study we report here, we directly test our hypotheses derived from the qualitative study about how the robot's behaviors contributed to where credit and blame were placed.

We examine credit and blame because they are critical to effective collaboration and decision making. If people assume too much personal responsibility for a task, it can lead to frustration and rigidity [2]. If, however, they assume too little responsibility or erroneously blame others, errors and conflict can result. Credit and blame are also central to our ideas about robots and morality. There has been much research on people assuming computers to have responsibility for ethical issues. Friedman argues that computers cannot have moral responsibility because they lack intentionality [3]. Nonetheless, through an experimental study, she found that most people actually attributed responsibility to computers. Results showed that 79% of the participants said that computers have decision–making capabilities and 45% of the participants judged computers to have intentions [4]. The study we report here explores the role of autonomy and transparency in attributions of credit and blame in human-robot interaction.

II. Theory and Hypotheses

A. Autonomy

Autonomy refers to "the degree to which team members experience substantial freedom, independence and discretion in their work" [5]. Tambe et al. observed that most robots are

Manuscript received March 15, 2006. This study was funded by NSF Grant IIS-0121426 to the second author. We would like to thank Christian Karega, Edwin Smolski and Iveshu Bhatia for running the experiments and all students who participated in the study.

Taemie Kim was with Stanford University, Stanford, CA 94305. She is now with The Media Lab at Massachusetts Instiutute of Technology, Cambridge, MA, 02139 USA (650-353-0100; e-mail: taemie@mit.edu).

Pamela Hinds is with the Department of Management Science & Engineering at Stanford University, Stanford, CA 94305 USA (e-mail: phinds@stanford.edu).

either "autonomous" or "non-autonomous" [6]. There is also a concept of "adjustable autonomy" where the level of self-sufficiency is variable depending on the situation [7]. For the purposes of our study, we focus on two levels of autonomy: (1) high autonomy with little need of human intervention and (2) low autonomy with need of constant human intervention.

From the Community Hospital experience, we noticed that when things went wrong or unexpectedly, many of the nurses blamed the robot even in cases when the fault was clearly their own or that of other coworkers. The existence of the robot seemed to have enabled a guilt-free direction of blame. From our analysis of the data at Community Hospital, we posit that individuals are more likely to attribute responsibility to the robot when they perceive the robot to be autonomous. In the process of decision making people expect mistakes. Because an autonomous robot appears to be exhibiting intention (by making judgments), we anticipate that people will assume it can make mistakes as well as be deserving of credit when its decisions have a positive outcome. This line of reasoning is consistent with work suggesting that a more human-like robot will attract more credit and blame than a machine-like robot [8], perhaps because people see these robots as more agent-like.

> ***Hypothesis 1.*** *When robots are more autonomous, individuals will attribute more credit and blame to the robots.*
>
> ***Hypothesis 2.*** *When robots are more autonomous, individuals will attribute less credit and blame to themselves.*
>
> ***Hypothesis 3.*** *When robots are more autonomous, individuals will attribute less credit and blame to other people who are also working with the robot.*

B. Transparency

We define transparency as the robot offering explanations of its actions. Research on attribution theory [9, 10] indicates that when people have more information, they are less likely to erroneously attribute blame to others. We speculate that providing explanations of a robot's actions, particularly ambiguous actions, will lead people to feel that they better understand the robot and to more accurate attributions about who is to blame for errors.

Sinha et al. defines transparency in a recommender system as "user understanding of why a particular recommendation was made" [11]. They showed for recommender systems that, in general, users prefer recommendations they perceive as transparent and feel more confident using the system. Especially for new items, users prefer transparent recommendations to non-transparent ones. Even for items that they already liked, users wanted to know why an item was recommended. This suggests that users want a justification of the system's decisions.

Transparency has effects other than causing people to like the system. Herlocker et al. presented experimental evidence showing that explanations can improve the acceptance of automated collaborative filtering (ACF) systems [12]. They first categorized the sources of error for ACF systems as model/process errors and data errors. By providing explanations for these errors, it allowed users a mechanism for handling errors associated with a recommendation.

A typical mobile robot does not provide direct and immediate feedback [13]**.** This causes the problem of delay in assigning appropriate credit and blame. A user cannot make proper decisions about when and how to use an agent unless the user can understand what the agent can do and how it will make decisions [14]. Further, they may have difficulty making the correct attributions in the absence of this information.

In Community Hospital, the nurses were constantly searching for reasons why the robot acted as it did. They would ask themselves and others, "What is going on here? Is the robot supposed to do this or did I do something wrong?" The low level of transparency led people to question even normal behaviors of the robot, sometimes leading people to think of correct behaviors as errors. This ambiguity resulted in incorrect attributions of credit and blame.

> ***Hypothesis 4.*** *When robots are more transparent, individuals are less likely to attribute credit and blame to the robots.*
>
> ***Hypothesis 5.*** *When robots are more transparent, individuals are less likely to attribute credit and blame to themselves.*
>
> ***Hypothesis 6.*** *When robots are more transparent, individuals are less likely to attribute credit and blame to other participants.*

C. Interaction between autonomy and transparency

Norman argued that high autonomy can sometimes be overwhelming and annoying to users because they feel a lack of control [15]. Transparency can decrease the level of annoyance because it lets people know what is happening so that they can direct the blame in the right direction.

We believe that a higher level of transparency in the robot deployed at Community Hospital may have improved workers' response to and acceptance of the robot. When interacting with a high autonomy robot, transparency can help users make sense of and develop a clearer understanding of the situation. However, when interacting with a low autonomy robot, we predict that transparency is unnecessary or even negative because the robot's behaviors are seen as less in need of explanation. Explanations may even be seen as distracting or inefficient. So, we predict that the effect for transparency (H4~6) will be stronger in the high autonomy case as compared with the low autonomy case.

> ***Hypothesis 7.*** *The effect of transparency is stronger when the robot is more autonomous.*

III. Method

We conducted a 2x2 laboratory experiment to test our hypotheses. The experiment was a between-subject design, manipulating autonomy and transparency of the robot. The robot was operated using a *Wizard of Oz* approach in which the robot was remotely controlled, seemingly autonomous. For consistency, a set of audio recordings were made of standard phrases said by the robot and played according to the condition.

A. Participants

We recruited 157 undergraduate and graduate students on a university campus to participate in a one-hour session and randomly assigned them to one of the four conditions. The mean age of participants was 20.14 and 53.5% of the participants were women. We collected no data on participants' ethnicity or national culture.

B. Tasks and procedures

Participants were brought into the lab in groups of four. During the session, we asked each participant to be in charge of one of the four part-stations of a toy manufacturing plant. Each part-station had toy pieces (such as Legos) and step-by-step instructions describing how to assemble those toy pieces into a structure. Each participant was asked to build three different assembly structures. These structures were to be individually delivered to another room by the robot.

The robot was introduced as a delivery robot that would visit each part-station every five minutes. The robot had a tray onto which participants could place their assemblies. The robot visited one part-station at a time. When approaching a part-station it announced, "Please place assembly number 101 on my tray." If the participant in the part-station was not ready with the assembly, s/he was instructed to say "Come back later." If the participant was ready s/he put the assembly on the robot's tray and the robot went to the next part-station. The next time the robot visited the same station, the robot asked for the next assembly in the sequence.

Participants were asked to fill out a brief demographic survey before the task and a post-task survey with questions about their experience after the task.

C. Manipulations

Autonomy had two levels for this experiment – low autonomy and high autonomy. For the high autonomy case, the robot made decisions about the status of the assembly and when to leave for the next station. When the robot accepted the assembly structure put on its tray it said, "This part is suitable for assembly. I have registered the part. I am leaving for the next station." When the robot rejected the assembly structure put on its tray it said, "This part is not suitable for assembly. Please remove it from my tray." And after removal it said, "I am leaving for the next station." There were two preplanned rejections where assemblies were intentionally designed to be loose. Preplanned rejections were inserted so that all participants could understand that the robot had the ability to reject parts. The robot also rejected assemblies that were clearly incorrect.

In the low autonomy conditions, the robot did not make any judgment about the assembly or when to leave. When a participant placed an assembly on the robot's tray the robot said "You have selected a part for assembly". The robot waited until the participant said "Robot, go" and then left for the next station. The participants were instructed differently according to the conditions of the session.

To check our autonomy manipulation, we asked participants three questions about the extent to which they thought that the robot had the ability to make task specific decisions (α = 0.72). Each question had a 7-point scale ranging from 1 (strongly disagree) to 7 (strongly agree). The results confirmed that those who worked with the high autonomy robot rated the robot as making more decisions (M=4.83, SD=1.16), than did participants working with the low autonomy robot (M= 2.99, SD=1.49). An analysis of variance (ANOVA) shows a statistically significance difference between the ratings, F(1, 157) = 74.57, p<.001.

Transparency also had two levels – low transparency and high transparency. In both cases the robot showed unexpected behavior during the third round of visits - it suddenly spun three times in one place. For the high transparency conditions, after the unexpected behavior the robot explained the reason for its action by announcing "I have recalibrated my sensors." For the low transparency conditions the robot offered no explanation. Our manipulation of transparency was explicitly associated with a behavior that was separate from the task to avoid potential confounds between autonomy and

TABLE I
CRONBACH'S ALPHA VALUE FOR THE DEPENDENT VARIABLES

SCALES	α
Attribution of blame to robot - The robot was responsible for any errors that were made in the task - The robot was to blame for most of the problems that were encountered in accomplishing this task	0.808
Attribution of credit to robot - Success on this task was largely due to the things the robot said or did - The robot should get credit for most of what was accomplished on this task.	0.678
Attribution of blame to self - I was responsible for any errors that were made in this task - I was to blame for most of the problems that were encountered in accomplishing this task	0.833
Attribution of credit to self - The success on this task was largely due to the things I said or did - I should get credit for most of what was accomplished on this task	0.852
Attribution of blame to other - Other participants were responsible for any errors that were made in this task - Other participants were to blame for most of the problems that were encountered in accomplishing this task	0.861
Attribution of credit to other - The success on this task was largely due to the things other participants said or did - Other participants should get credit for most of what was accomplished on this task	0.781

Note. N=157. Cronbach's Alpha is a measure of the reliability of the scale as a whole. Alpha ranges from zero to 1.0 (highest).

transparency.

D. *Measures*

Our six dependent variables were the level of credit and blame attributed to the robot, to oneself, and to other participants in the 4-person team. These were measured through questions on the post-task survey. All questions were answered on a 7-point scale ranging from 1 (strongly disagree) to 7 (strongly agree). For each dependent variable, we asked participants two questions and averaged the two values to create a single measure. Table 1 shows the Cronbach's α for each scale.

IV. Results

A. *Effects of autonomy*

In hypotheses H1-H3, we argued that autonomy would lead to more attributions of responsibility (credit and blame) to the robots and less to oneself and to other team members. Our data provide good support for this. We found that participants attributed more blame to the high autonomy robot (M=2.96, SD=1.19) than the low autonomy robot (M=2.18, SD=1.49), and the difference was significant in an ANOVA with autonomy and transparency as factors, F(3,153)=13.32, p<.001. There was, however, no significant difference for credit to the robot (M= 2.75, SD=1.32, M=2.54, SD=1.31, respectively).

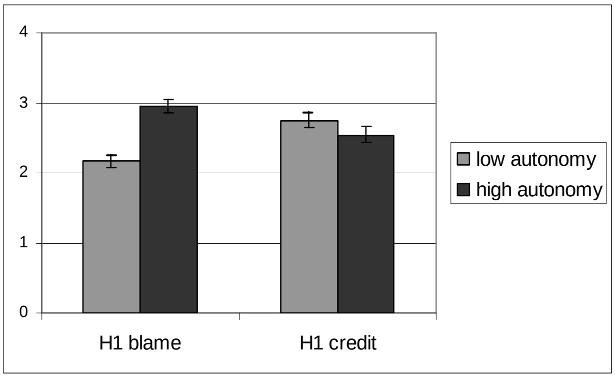

Fig. 1. The effect of autonomy on attribution of blame and credit to the robot (H1)

Similarly, our results show that participants attributed significantly less blame to themselves for errors that occurred in the task when they worked with the high autonomy robot (M=3.87, SD=1.61) than when they worked with the low autonomy robot (M=4.72, SD=1.57), F(3,153)=11.53, p<.001. The participants also thought they should get less credit when working with the high autonomy robot (M= 4.49, SD=1.54) than the low autonomy robot (M=4.80, SD=1.50), but the difference was not significant, F(3,153)=1.64, p=.21.

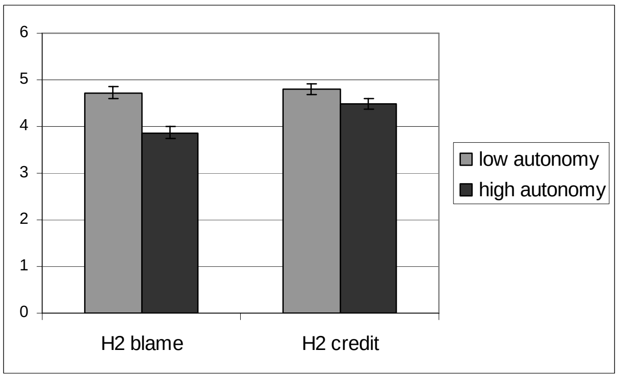

Fig. 2. The effect of autonomy on attribution of blame and credit to oneself (H2)

Finally, participants attributed significantly less blame to the other participants when working with the high autonomy robot (M=3.28, SD=1.43) than with the low autonomy robot (M=3.97, SD=1.50), F(3,153)=8.97, p<.05. They also attributed significantly less credit to other participants when they worked with high autonomy robot (M= 4.06, SD=1.50) than when they worked with the low autonomy robot (M=4.50, SD=1.42), F(3,153)=3.96, p<.05.

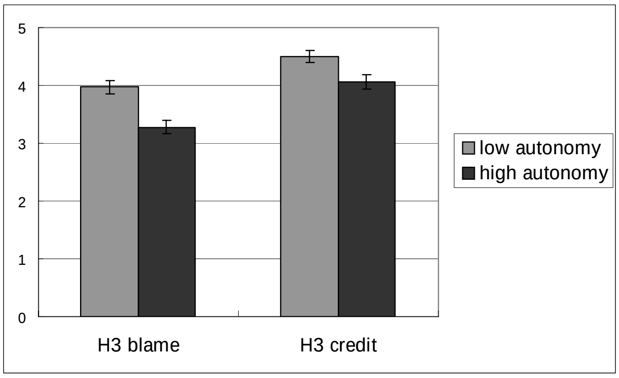

Fig. 3. The effect of autonomy on attribution of blame and credit to other participants (H3)

B. *Effects of transparency*

In our second set of hypotheses (H4-H6), we argued that transparency would lead to fewer attributions of blame to anyone, including the robot. We found, however, only moderate support for these hypotheses. There was little difference in attributions of credit or blame to the robot (H4) in the high transparency as compared with the low transparency conditions (M=2.71 vs. 2.43 for blame and M=2.72 vs. 2.57 for credit). Similarly, when examining the attribution of credit and blame toward oneself (H5), there was little difference when working with the high transparency robot as compared with the low transparency robot (M=4.11 vs. 4.48 for blame and M=4.64 vs. 4.65 for credit).

In support of hypothesis 6, however, participants attributed significantly less blame to other participants when they worked with the high transparency robot (M=3.33, SD=1.36) as compared to when they worked with the low transparency robot (M=3.91, SD=1.58), F(3.153)=6.45, p<.05. They also attributed less credit to other participants when they worked with the high transparency robot (M=3.90, SD=1.37) than when they worked with the low transparency robot (M=4.64, SD=1.48), F(3,153)=10.81, p<.001.

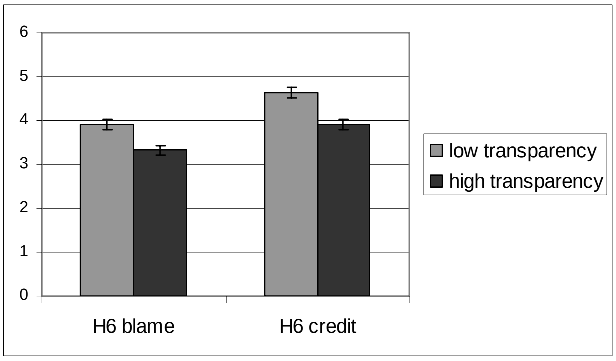

Fig. 4. The effect of transparency on attribution of blame and credit to other participants (H6)

C. Interaction between autonomy and transparency

In hypothesis 7, we argued that transparency would have a stronger effect when the robot was more autonomous. Our reasoning was that autonomy can make the robot's actions less clear, so transparency is particularly important to help explain these actions. To test the hypothesis, we compared the effect of transparency on our six dependent variables in the high autonomy case and the low autonomy case. As predicted, the results showed that transparency had a much larger effect on credit toward other participants in the high autonomy conditions (M=3.48, SD=1.32 for the high transparency case and M=4.61, SD=1.47 for the low transparency case) than in the low autonomy conditions (M=4.32, SD=1.31 for the high transparency case and M=4.67, SD=1.51 for the low transparency case). A two-way ANOVA analysis with autonomy and transparency as factors shows a marginally significant effect in the expected direction for the attribution of credit to other participants, F(3,153)=2.94, p<0.10. The effect of transparency on credit to other participants is much stronger when working with the high autonomy robot than the low autonomy robot (Fig. 5). The interaction effect was not significant for any other dependent variables.

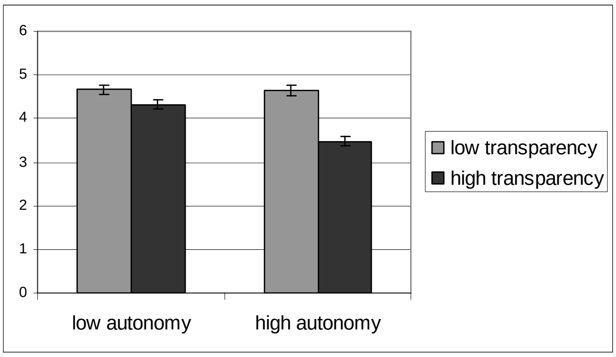

Fig. 5. The interaction effect of autonomy and transparency on attribution of credit on other participants (H7)

V. Discussion

Our results suggest that when a robot has more autonomy, people will attribute more blame to the robot and less to themselves and their co-workers. This is consistent with our prediction that autonomy will contribute to a shift in responsibility from the person to the robot. It is interesting to note, however, that attributions of credit did not show the same pattern. That is, people shifted blame for errors, but not credit for successes to the robot. These findings have several implications. First, it appears that people begin to abdicate responsibility for errors when faced with autonomous robots. This may reduce rigidity, particularly in high threat situations, but it also could reduce peoples' conscientiousness in the task. Thus, level of autonomy may be an important design consideration that depends on the desired level of human responsibility in the particular situation.

Our hypotheses regarding transparency were only partially supported. A more transparent robot, one that explained its unexpected behavior, did not significantly affect the credit or blame participants attributed to oneself. However, it had a marginally significant effect on the credit attributed to other participants.

The effects for attributions toward the robot, though insignificant, were in the expected direction. This suggests that by explaining its actions, the robot has allowed the participants to attribute less responsibility to other users while shifting that blame slightly toward the robot. This finding is consistent with what we observed in Community Hospital. When workers noticed inexplicable behavior or errors by the robot, they often blamed co-workers for having done something to "mess up" the robot. Consistent with attribution theory, people tend to blame others for errors more than they blame themselves [10]. When information is provided to explain behaviors, erroneous attributions of blame are tempered [9].

Our manipulation for transparency involved having the robot either explain or not explain an unexpected behavior. We expected that this would increase participants' perception that they understood the robots' behavior. We were surprised, however, to find that the relationship between transparency and participants' self-reported understanding of the robot was negative. When we created a scale for two questions (α=0.80) about how much participants understood the reasons behind the robot's action, the means were M=3.94, SD=1.68 for the high transparency robot and M=4.51, SD=1.61 for the low transparency robot, and the difference was significant, F(1,157)=4.73, p<.05. Thus, transparency was associated with *less*, not more understanding. Further investigation suggests that this is likely a result of our transparency manipulation. In the high transparency conditions, the robot explained that it was recalibrating its sensors. The participants in our study may not have known what this meant and were therefore further confused by the robot's explanation. Consistent with this, transparency and understanding were correlated with the participants' primary discipline of study, F(1,157)=4.00, p<0.05. Students in non-technical majors reported understanding the robot less in the high vs. low transparency conditions (M=3.52, SD=1.61 vs. M=4.46, SD=1.61) whereas students in technical majors did not (M=4.56, SD=1.60 vs. M=4.62, SD=1.61). This suggests that the effect of transparency is highly dependent on the match between the robots' explanation of its actions and the background knowledge of participants. That is, if the robot explains its behavior in a language not suited to the user, transparency can create more rather than less confusion.

Although our effects were somewhat inconclusive for the interaction between autonomy and transparency, we believe they provide some evidence for the value of transparency for autonomous robots. As can be seen from Fig. 5, transparency had little effect on credit to other participants when the robot was low in autonomy, but when the robot was high in autonomy, transparency had a significant effect on reducing the credit attributed to other participants. These findings suggest that when a robot explains its actions, particularly actions that are ambiguous, it may lead people to more accurately attribute credit (and perhaps blame). Therefore transparency should be considered when designing autonomous robots.

This study has examined the effect of the robot's behaviors in a collaborative group task and provides possible design insights for autonomous mobile robots. Continued work in this area will improve the likelihood of robots being accepted as group members and attributed with the appropriate credit and blame for a given situation.

REFERENCES

[1] United Nations and the International Federation of Robotics, World Robotics 2004, United Nations, New York, 2004

[2] N.C. Roberts and L. Wargo, "The Dilemma of Planning in Large-Scale Public Organizations: The Case of the United States Navy", Journal of Public Administration Research and Theory, 1994, pp. 469-491.

[3] Friedman, B., Moral Responsibility and Computer Technology, Erin Document Reproduction Services, April 1990.

[4] B., Friedman and L., Millett, "It's the computer's fault: reasoning about computers as moral agents", Proceedings of the CHI 1995, Conference on Human Factors on Computer Systems, ACM, New York, 1995.

[5] B., Kirkman, B., Rosen, P., Tesluk and C.B. Gibson, "The impact of team empowerment on virtual team performance", Academy of Management Journal, 2004, pp. 58-74

[6] M., Tambe, D., Pynadath, and P. Scerri, "Adjustable Autonomy: A Response", Intelligent Agents VII Proceedings of the International workshop on Agents, theories, architectures and languages, 2001

[7] D., Perzanowski, A., Schultz, E., Marsh and W., Adams, "Two ingredients for my dinner with R2D2: Integration and Adjustable Autonomy". Papers from the 2000 AAAI Spring Symposium Series, AAAI Press, Menlo Park, CA, 2000, pp. 1-6

[8] P., Hinds, T., Roberts and H., Jones, "Whose job is it anyway? A study of Human-Robot Interaction in a Collaborative Task", Human-Computer Interaction, Vol. 19, 2004, pp.151-181,

[9] E., Jones and R., Nisbett, "The Actor and the Observer: Divergent perceptions of the causes of behavior", In. E. Jones, D. Kanouse, H. Kelley, R. Nisbett, S. Valins, and B. Weiner (eds.), Attribution: Perceiving the Causes of Behavior, General Learning Press, Morristown, NJ, 1972, pp 79-94.

[10] L., Ross, "The Intuitive Psychologist and His Shortcomings: Distortions in the Attribution Process", Advances in Experimental Social Psychology, 10, 1977, pp. 174-220.

[11] R., Sinha and K., Swearingen, "The Role of Transparency in Recommender Systems", Proceedings of CHI'2002, Conference on Human Factors on Computer Systems, ACM, New York, 2002

[12] J., Herlocker, J.A., Konstan, and J., Riedl, "Explaining Collaborative Filtering Recommendations". Proceedings of 2000 ACM Conference on Computer Supported Cooperative Work, Philadelphia, PA, 2000, pp. 241-250.

[13] M., Matari, "Reinforcement Learning in the Multi-Robot Domain", Autonomous Robots vol. 4, 1997, pp. 73-83

[14] C., Heckman, and J., Wobbrock, "Liability for autonomous agent design", Proceedings of the second international conference on Autonomous agents, Minneapolis, Minnesota, United States, 1998, pp.392-399.

[15] D., Norman, "How might people interact with agents?" Com*munications of the ACM 37(7)*, 1994, pp. 68-71.