Social Responses to Collaborator: Dilemma Game with Human and Computer Agent

Kazuhisa Miwa (miwa@is.nagoya-u.ac.jp)
Graduate School of Information Science, Nagoya University
Nagoya, 464-8601 JAPAN
Hitoshi Terai (terai@sie.dendai.ac.jp)
School of Information Environment, Tokyo Denki University
Inzai, Chiba, 270-1355 JAPAN
Satoshi Hirose (hirose@cog.human.nagoya-u.ac.jp)
Graduate School of Information Science, Nagoya University
Nagoya, 464-8601 JAPAN

Abstract

We performed an experiment where participants engaged in the Prisoner's Dilemma game either with a human or a computer agent. In the experiment, we controlled two factors: (1) expectation about a partner, i.e., whether a partner is believed to be a human or computer agent, and (2) actual partner's behavior, i.e., whether (a) a partner performs with human-like sophisticated behavior or (b) simple mechanical behavior. Participant decision-making behavior showed that their defect actions greatly increased when instructed that their partner was a computer agent; the effect of the actual partner's behavior was limited. Personality impression tests showed that the partner's individual desirability correlated to the number of times defected by the partner, but the partner's social desirability correlated to the number of defect actions that the participants offered. Our conclusion is that humans actually generate social relationships with computer agents, as the Media Equation studies have insisted; however these relationships are relatively different from those with humans.

Keywords: Human-Human Interaction, Human-Agent Interaction; Prisoner's Dilemma.

Introduction

Interaction is a crucial research topic in cognitive science. Recently, since various types of artifacts have entered our society, many researchers have begun to show interest not only in human and human interaction (HHI) but also in human and computer-agent interaction (HAI).

One research paradigm emerging at the intersection of HHI and HAI studies is the Media Equation framework. Media Equation studies have revealed that human beings often relate to computers as they do to other human beings. This tendency has been widely confirmed from the following viewpoints: politeness (Nass, , Moon, & Carney, 1999), reciprocity (Fogg & Nass, 1997), personality (Moon & Nass, 1996), in-group/out-group differences (Nass, Fogg, & Moon, 1996), and ethnicity (Nass, Lsbister, & Lee, 2001). This research paradigm investigated human subject responses to computers in typical social situations, and confirmed that their responses to computers resembled responses to humans observed in social-psychological experiments.

Media Equation studies have successfully delineated the important aspects of human interaction with humans and computer agents. However, to investigate the interaction more carefully, we need a new experimental paradigm that controls the factor of partner as an independent variable. In the experimental paradigm, subject behavior with computer agents is directly compared to the behavior with humans, where two experimental situations, interacting with a human and interacting with a computer agent, are set up.

There are related studies. In one experiment, subjects played Monopoly with humans or computers, and experienced higher levels of aggressive feelings after playing with computers than with humans (Williams & Clippinger, 2002). In other experiments, subjects engaged in the Prisoner's Dilemma game. Their behavior largely varied depending not only on difference of a human and computer partner but also on the appearance of a partner displayed on a monitor (Kiesler, Waters, & Sproull, 1996; Parise, et al., 1999). These studies suggest that humans may actually generate social relationships with computer agents; however such relationships are relatively different from those with other humans.

Note here that the independent variable, an interaction partner, is divided into two factors: (1) expectation about a partner, i.e., whether a partner is believed to be a human or computer agent, and (2) actual partner's behavior, i.e., whether (a) a partner performs with human-like sophisticated behavior or (b) with simple mechanical behavior.

These two factors should be considered independently. Actually, studies on human computer-agent interaction indicate that if a serious gap between expectations from appearance of artifact and actual function performed by the artifact emerges, it often becomes impossible for people to interact with the artifact. For example, during interaction with a human-like humanoid robot, people have noticed its lack of sophisticated functions; and disappointedly terminated interaction. Another example is that people can converse with a very simple chat program such as Eliza (Weizenbaum, 1996), but in trials to create humanoid robots whose appearance very closely resembles humans', the "uncanny valley" emerges and interaction breaks down (see a CogSci2005 workshop: http://www.androidscience.com/).

The above cases indicate that two factors, expectation about a partner and the actual partner's behavior, are crucial to determine the quality of HHI and HAI. Based on this insight, we previously proposed an experimental approach

called the Illusion Experiment Paradigm where the two factors can be independently controlled. We experimentally investigated the natures of HHI and HAI using a simple problem-solving task. In the current study, we report experimental results using the Prisoner's Dilemma game where deeper social interaction is needed to perform the task.

Task

The Prisoner's Dilemma game is an abstraction of social situations where each participant faces two alternative actions: (1) cooperation: i.e., doing the socially responsible thing and (2) defection: i.e., acting based on self-interest regardless how this might harm the partner. Each participant is better off defecting regardless of the partner's choice, but the sum of the participants' payoffs is maximized if both participants choose to cooperate; so a dilemma emerges. Table 1 shows the payoff matrix used in the current study. For example, when both subjects offer cooperation, both receive 120 Yen; however when one subject offers cooperation while the other offers defection, the former receives nothing while the latter receive 180 Yen. They were instructed that they would be actually paid based on the payoff matrix.

Table 1: Payoff matrix in Prisoner's Dilemma game.

	B: cooperate	B: defect
A: cooperate	A: 120 Yen/B: 120 Yen	A: 0 Yen/B: 180 Yen
A: defect	A: 180 Yen/B: 0 Yen	A: 60 Yen/B:60 Yen

Illusion Experiment Paradigm

In this study, we independently controlled two factors: (1) expectation about a partner and (2) the actual partner's behavior by developing the Illusion Experiment Paradigm (see Figure 1).

Actual partner's behavior

As an independent variable, the first experimental factor is related to the actual partner's behavior. This factor was controlled by manipulating the partner with which subjects actually collaborated. Three cases were set up: (1) collaboration with a human subject (w/ Human), and (2) collaboration with a computer agent. The former represents a case where a partner computer agent is sophisticatedly designed so that it behaves almost identical to humans. The latter case was subdivided into two sub cases: (2a) collaboration with an agent who uses the cooperation strategy in decision-making (w/ C-agent), and (2b) collaboration with an agent who uses the defection strategy (w/ D-agent). The C-agent offered ten cooperate actions and only one defect action in the 6th trial among eleven trials. The D-agent offered six cooperate actions and five defect actions in the 2nd, 4th, 5th, 7th, and 10th trials.

The first factor was manipulated as follows. When collaborating with a human subject, each terminal was connected to the Internet through LAN, and each subject solves the task with a partner who simultaneously engages in this task using another computer terminal connected through the Internet. On the other hand, when collaborating with a computer agent, each terminal operates independently from the others, and each subject solves the task with an agent established on a computer.

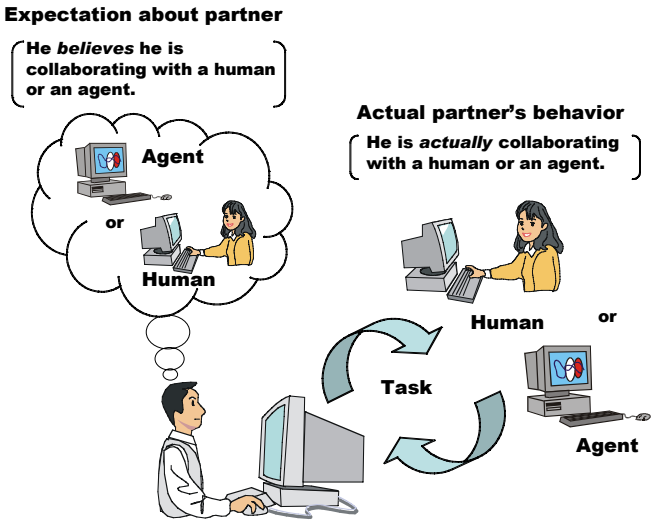


Figure 1: Illustration of Illusion Experiment Paradigm.

Expectation about a partner

The second factor is related to expectation about a partner caused by the experimenter's instructions. This factor was controlled by manipulating the subject with whom they believed they were collaborating. Two cases were set up: (1) a case where subjects were instructed to collaborate with a program installed on the computer they were manipulating, and (2) a case where they were instructed to collaborate with a human subject sitting in front of another computer terminal, communicating by the Internet.

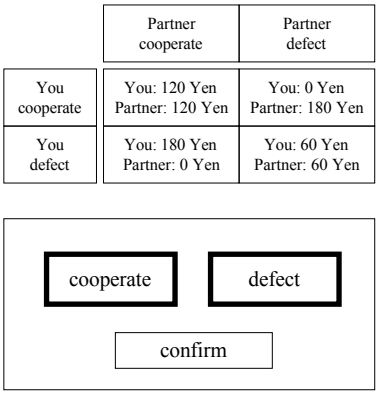
When introducing subjects to a collaboration situation with a human subject (Human instruction), in the initial stage of the experiment, subjects introduced themselves to their partners in face-to-face situations, and then moved to their respective computer terminals. On the other hand, when collaboration with a computer agent was given (Agent instruction), subjects sat in front of an assigned computer terminal and engaged in the task, believing they were working with a computer program established on their computer.

Experiment

Procedure

Subjects continuously made eleven decisions one by one. Prior to each decision, they were not informed of the partner's decision. After being informed of the partner's

decision in the preceding trial, then they were required to make their next decision. Subjects engaged in the task using a computer terminal. Figure 2 shows an example screen shot of the computer terminal used in the experiment.



	you	partner
1	120	120
2	0	180
3	0	180
4		
5		
6		
7		
8		
9		
10		
11		
sum	120	480

Figure 2: Example screen shot of computer terminal for experiment.

Questionnaires

After all eleven decisions were completed, a questionnaire, developed by Hayashi (1978), was performed to determine the personality impressions of the partner. The questionnaire consists of 20 pairs of adjectives translated from Japanese as follows: (1) active/passive, (2) bad-natured/good-natured, (3) rude/respectful, (4) friendly/unfriendly, (5) lovable/hateful, (6) generous/stingy, (7) reclusive/sociable, (8) responsible/undependable, (9) tidy/careless, (10) unblushing/bashful, (11) profound/shallow, (12) exhilarating/gloomy, (13) grand/subservient, (14) pleasant/unpleasant, (15) prudent/imprudent, (16) approachable/distant, (17) ambitious/lazy, (18) confident/diffident, (19) anxious/calm, and (20) cruel/kind. Subjects estimated the personality impressions of their partner using a 1 to 7 scale where 7 is the maximum (active) and 1 is the minimum (passive).

Participants

One hundred forty-five undergraduates participated in the experiment. They were randomly assigned to one of the six conditions, where the number of subjects in each condition was as equal as possible. As a result, the numbers of subjects in the Human (instruction)/Human (actual partner), Human/C-agent, Human/D-agent, Agent/Human, Agent/C-agent, and Agent/D-agent were 24, 26, 25, 26, 23, and 21, respectively.

Results

Decision-making behavior

Figure 3 shows the subjects' decision-making behavior where the vertical axis indicates the rate of the subjects' defect actions, and the horizontal axis indicates each experimental condition.

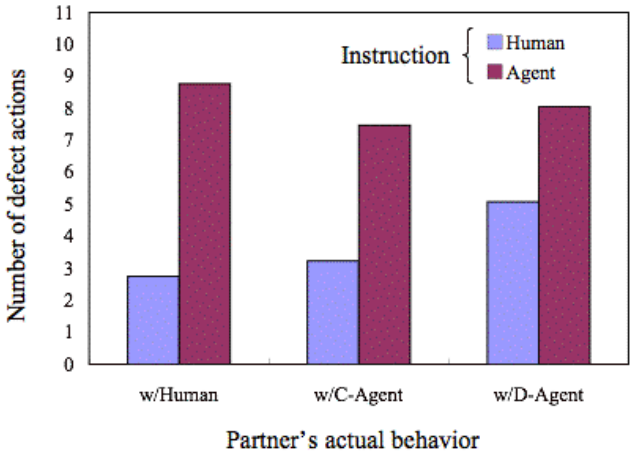


Figure 3: Number of defect actions in participants' decision-making.

A two (instruction) x three (behavior) ANOVA revealed that the main effect of instruction reached significance ($F(1,139)=89.927$, $p<0.01$), but not the main effect of the actual behavior ($F(2,139)=2.22786$, n.s.). The interaction also reached significance ($F(2, 139)=3.61$, $p<.05$). The simple main effect of the actual behavior at Human instruction revealed significance ($p<.05$), where a Ryan' multiple comparison analysis showed that the rate of defection was higher at w/D-agent than at w/Human or w/C-agent ($p<.01$ and $p<.05$). Otherwise the simple main effect of the actual behavior at Agent instruction did not reach significance.

The analysis shows that people offered more defect actions when instructed that their partner was a computer agent than a human. When instructed that their partner was a computer agent, the rate of defect actions was consistently large without depending on their partner's behavior. On the contrary, when instructed that their partner was a human, they offered more defect actions when collaborating with a partner who offered more defect actions than when collaborating with a partner who offered fewer defect actions.

Estimation of impressions of a partner

We focused on three factors extracted by Hayashi (1978) for analyzing personality impressions measured in the questionnaires: activity, social desirability, and individual desirability. Nine traits have higher factor loadings for

activity: active/passive, bad-natured/good-natured, rude/respectful, reclusive/sociable, unblushing/bashful, ambitious/lazy, confident/diffident, anxious/calm, and cruel/kind. We averaged the nine scores for a representative activity score. Similarly, the average score for social desirability was calculated from five items: responsible/undependable, tidy/careless, profound/shallow, grand/subservient, prudent/imprudent, and the average score for individual desirability from lovable/hateful and friendly/unfriendly.

Figures 4(a) to (c) indicate the average scores for activity, social desirability, and individual desirability estimated in each of the six conditions. In activity, a two (instruction) x three (behavior) ANOVA revealed that the main effect of the actual behavior reached significance ($F(2, 132)=15.44$, $p<.01$), whereas neither the main effect of instruction nor interaction reached significance ($F<1$, n.s.; $F(2, 132)=1.33$, n.s.). A Ryan' multiple comparison analysis showed that the scores at w/Human and w/D-agent were higher than that at w/C-agent ($p<.01$; $p<.01$).

In social desirability, the same ANOVA revealed that the main effect of instruction reached significance ($F(1,132)=23.900$, $p<0.01$), while the main effect of the actual behavior did not ($F(2,132)=2.912$, n.s.). The interaction also reached significance ($F(2, 139)=6.06$, $p<.01$). The simple main effect of the actual behavior at Human instruction revealed significance ($p<.05$), where a Ryan' multiple comparison analysis showed the score was higher at w/Human and w/C-agent than that at w/D-agent ($p<.01$ and $p<.05$). Otherwise the simple main effect of the actual behavior at Agent instruction did not reach significance.

In individual desirability, the same ANOVA detected the main effects of both instruction and actual behavior ($F(1,132=17.566, p<0.01$; $F(2,132)=11.722, p<0.01$). The interaction also reached significance ($F(2, 139)=10.00$, $p<.01$). Both the simple main effects of the actual behavior at Human instruction and at Agent instruction revealed significance ($p<.01$; $p<.01$). A Ryan' multiple comparison analysis showed that in Human instruction the scores were higher at w/Human and w/C-agent than that at w/D-agent, and in Agent instruction the score was higher at w/C-agent than those at w/Human and w/D-agent.

Overall results showed that impressions of a partner were influenced by both expectation about a partner and the actual partner's behavior. This finding will be examined in the following discussion and conclusions.

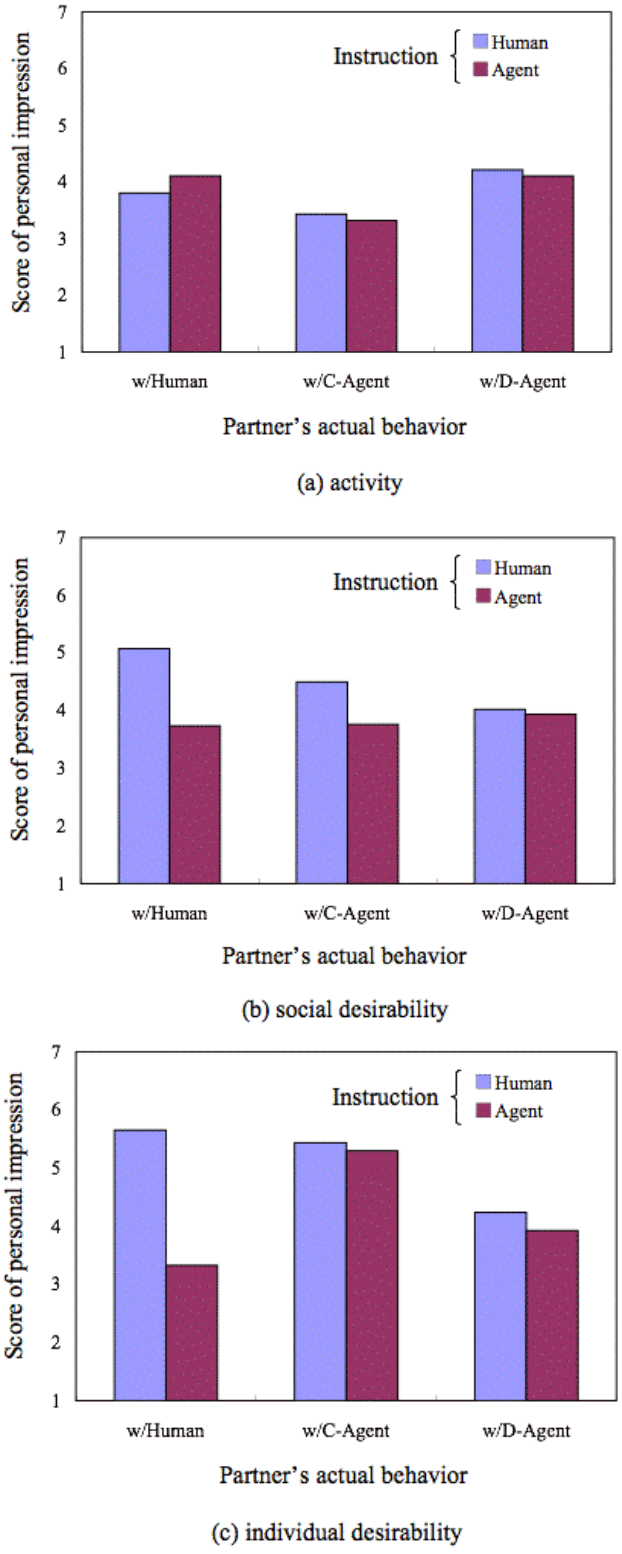


Figure 4: Personal impressions of a partner.

Discussion and Conclusion

Sociality and difference in responses to a partner

As shown in Figure 3, subject decision-making was mainly influenced by the factor of instruction rather than the partner's actual behavior.

The experiment in Miwa & Terai (2006) was based on the same research paradigm, i.e., the Illusion Experiment Paradigm, using a rule discovery task, Wason's 2-4-6 task,

which is widely used in laboratory studies on human scientific discovery. The first experimental result showed that problem-solving behavior, such as hypothesis-testing strategies and reference to partner's hypotheses, is largely influenced by the actual partner's behavior but not by instruction.

On the other hand, the second experimental result showed that in reciprocity behavior (i.e., people giving much information to a partner when receiving much information from a partner), the main effect of instruction did not appear but the interaction of instruction and amount of received information did. On the contrary, in the current experiment with the Prisoner's Dilemma game, a strong main effect of instruction emerged in subjects' decision-making behavior.

We expected the degree of sociality behind subject behaviors to increase in the order of problem-solving behavior, reciprocity behavior, and decision-making behavior focused on in this study. The above results indicate that the differences in responses to humans and computer agents are becoming salient in situations where sociality is largely required.

Difference in responses to humans and computers

Our experiment indicated strong evidence of the effect of instruction, i.e., people performing different decisions depending on whether they expect that their partner is a human or computer agent, even though the partner's actual behaviors are identical. This result is interesting when compared to the findings in the current study with the experimental results obtained in the preceding social-psychological experiments, which have consistently indicated human robust orientation to cooperative behavior. For example, Kiyonari et al. performed a very realistic Prisoner's Dilemma experiment where subjects were actually handed real money calculated based on the game's payoff matrix (Kiyonari, Tanida, & Yamagishi, 2000). They compared the result of this experiment with the vignette experiment where subjects just gained game points that provided no practical benefits for their real lives. They confirmed that subjects offered more cooperate actions in the realistic than in the vignette situation. This finding suggests that people prefer cooperation to defection; it is a very strong and fundamental nature of human behavior when working with other humans. On the other hand, in our experiment, simple instruction that a partner is a computer agent had the subjects drastically shift their behavior to offer defect actions. This also indicates a critical difference in subject responses to humans and computers.

Effects of instruction

The effect of instruction confirmed in this study has also been found in other studies. First, Yamamoto et al. conducted an experiment where subjects played Shiritori, a popular word game in Japan, with a partner through a computer terminal (Yamamoto, et al., 1994). Shiritori is a game played by saying a word starts with the last syllable of the word given by the previous player. Subjects were actually playing Shiritori with a computer program installed in a computer. Instruction was controlled; in one situation subjects were informed that the partner was a computer program and in the other situation the partner was a human at another campus connected by the Internet. The degree of enjoyment of the game felt by the subjects depended on the instruction.

Second, in the field of the development of computer agents, the Wizard of Oz method has been established for collecting corpus data of natural conversations with computer agents (Dahlbak, Jonsson, & Ahrenberg, 1993). This situation is the reverse of the previous situation in the Shiritori experiment; in the Wizard of Oz method participants are instructed that their partner is a computer agent even though the partner is actually a human. This method, which has been widely accepted as valid, is based on the idea that this instruction can make a natural interaction setting with a computer agent even though the actual partner is a human.

Third, the effect of instruction has also been found in an experiment conducted in the context of Media Equation studies. Sundar & Nass performed an experiment to understand whether subjects attribute a partner to a computer agent itself or to a programmer who developed the computer agent when interacting with the agent (Sundar & Nass, 2000). They hypothesized that if the effect of instruction emerges between situations where subjects are informed that the partner is a computer agent and where the partner is the programmer, then they will admit that the subjects attribute their partner to the computer agent, not to the programmer. We are focusing on their second experiment. They performed the same experiment in one situation where the instruction revealed that the partner was a computer agent and in another situation where the instruction revealed the partner to be a human in another room connected by the Internet. This experiment was originally intended to investigate the nature of subject attribution; however, the result also confirmed the effect of instruction.

The above results have consistently confirmed the effect of instruction indicating that people generate different relationships with a partner depending on whether they expect that the partner is human or a computer agent.

Media Equation perspective

The above results provide a slightly different perspective on HHI and HAI from the findings confirmed by Media Equation studies. On the other hand, the results of the questionnaire in which we obtained the data of subject estimations of their impressions of a partner provided further evidence indicating the similarity of human social responses to computers and other humans.

When comparing the rate of subject defect behavior shown in Figure 3 and the estimation of partner's social desirability shown in Figure 4(b), the pattern of scores in the two graphs indicates a reversal relationship. This suggests that as the number of defections that subjects offer is larger,

estimations of their partner's social desirability decrease. Additionally, we also confirm a correlation between the number of received defect actions from a partner and subjects estimations of the partner's individual desirability. The C-agent was designed to offer a defect once among eleven trials while the D-agent was designed to offer five defect actions. In cases of interaction with a human, the average number of defection was 2.75 times when instruction revealed the partner was a human and 8.77 times when instruction revealed the partner was a computer agent. When inspecting the subjects' estimation of their partner's individual desirability shown in Figure 4(c) while confirming these facts, we also hypothesize that as the number of times defected by a partner increases, the estimation of the partner's individual desirability will decrease.

To verify the above insight, we analyzed the correlation between the number of defections that the subjects offered/ the number of times defected by a partner and the subjects' estimations on the impressions of a partner. Table 2 shows the result. The number of defections considerably correlates to the degree of the partner's social desirability while the number of times defected by a partner correlates to the degree of the partner's individual desirability. These results imply that people feel much familiarity with a partner who performs cooperative behavior regardless whether the partner is a human or a computer agent. More interestingly, when people offer defect actions to a partner, they tend to reduce their estimations of their partner's social desirability. This behavior is considered a kind of reducing cognitive dissonance behavior that justifies their defect actions (Festinger, 1957). This suggests that people have to try to cancel their inner cognitive dissonance even though a partner is an artifact such as a computer agent.

Table 2: Correlation between numbers of defecting and defected and personality impression of partner.

	# defecting	# being defected
Social desirability	$r = -.49, p < .01$	$r = -.23, p < .01$
Individual desirability	$r = -.22, p < .05$	$r = -.59, p < .01$

References

Dahlbak, N., Jonsson, A., & Ahrenberg, L. (1993). Wizard of oz studies: why and how. Knowledge-based systems, 6, 258-266.

Festinger, L. (1957). A theory of cognitive dissonance. Standard University Press.

Fogg, B., & Nass, C. (1997). Do users reciprocate to computers? Proceedings of the CHI conference (Atlanta, GA). New York: Association of Computing Machinery.

Hayashi, F. (1978). The fundamental dimensions of interpersonal cognitive structure. Bulletin of School of Education, Nagoya University, 25, 233-247. (in Japanese)

Kiesler, S., Waters, K., & Sproull, L. (1996). A Prisonor's Dilemma Experiment on Cooperation With People and Human-Like Computers. Journal of Personality and Social Psychology, 70, 47-65.

Kiyonari, T., Tanida, S., & Yamagishi, T. (2000). Social exchange and reciprocity: confusion or a heuristic? Evolution and Human Behavior, 21, 411-427.

Miwa, K., & Terai, H. (2006) Analysis of Human-Human and Human-Computer Agent Interactions from the viewpoint of design of and attribution to a Partner. Proceedings of 28th annual meeting of the cognitive science society, 597-602.

Moon, Y. & Nass, C. (1996). How real are computer personalities? Psychological responses to personality types in human-computer interaction. Communication Research, 23, 651-674.

Nass, C., Fogg, B., & Moon, Y. (1996). Can computers be teammates? International Journal of Human-Computer Studies, 45, 669-678.

Nass, C., Moon, Y., & Carney, P. (1999). Are respondents polite to computers? Social desirability and direct responses to computers. Journal of Applied Social Psychology, 29, 1093-1110.

Nass, C., & Sundar, S. (2000). Source Orientation in Human-Computer Interaction: Programmer, Networker, or Indipendent Social Actor? Communication Research, 27, 683-703.

Nass, C., Lsbister, K., & Lee, E. (2001). Truth is beauty: researching embodied conversational agents. In Castells, J. (Ed.) Embodied conversational agents. CAMBRIDGE, MA: MIT Press.

Parise, S., Kiesler, S., & Sproull, L., & Waters, K. (1999). Cooperating with life-like interface agents. Computers in Human Behavior, 15, 123-142.

Sundar, S,. & Nass. C. (2000). Source orientation in human-computer interaction: Programmer, networker, or independent social actor? Communication Research, 27, 683-703.

Weizenbaum, J. (1996). A computer program for the study of natural language communication between man and machine. Communications of the Association for Computing Machinery, 9, 36-45.

Williams, R., & Clippinger, C. (2002). Aggression, competition and computer games: computer and human opponents. Computers in Human Behavior, 18, 495-506.

Yamamoto, Y., Matsui, T., Hiraki, K., Umeda, S., & Anzai, Y. (1994). Interaction with a computer system: a study of factors for pleasant interaction. Cognitive studies, 1, 107-120. (in Japanese)