Looking Real and Making Mistakes

Yuqiong Wang[1], Peter Khooshabeh[2,⋆], and Jonathan Gratch[2,⋆⋆]

[1] Rochester Institute of Technology, Rochester NY, 14623, USA
wang@mail.rit.edu
[2] University of Southern California, Institute for Creative Technologies
Playa Vista CA, 90094, USA
{khooshabeh,gratch}@ict.usc.edu

Abstract. What happens when a Virtual Human makes a mistake? In this study we investigate the impact of a VH's conversational mistakes in the context of persuasion. Users interacted with a VH that told persuasive information, and they were given the option to use the information to complete a problem-solving task. The VH occasionally made mistakes such as not responding, repeating the same answer, or giving an irrelevant reply. Results indicated that a VH who makes conversational mistakes is capable of social influence. Individual differences also shed light on the cognitive processes of users who interacted with error-prone VHs. We discuss the implications of these results with regard to VH design.

Keywords: Virtual Humans, Human-Virtual Human Interaction, Virtual Reality, Photorealism, Conversational Mistakes, Need For Cognition.

1 Introduction

To err is human, to forgive, divine.
– Alexander Pope

Virtual Humans (VHs) are representations of humans in virtual environments, in the form of online puppetry of actual humans (i.e., avatars), or computer algorithms simulating people (i.e., agents). The technology has made possible Human-Virtual Human (H-VH) interaction in various applications such as medicine [1], education [2], and therapy [3]. Limitations of this technology, on the other hand, have left VHs prone to errors. For instance, a VH empowered by natural language processing may fail to answer a user's question, or give a nonsensical response [4]. Based on theories that suggest people treat computers as social actors [5], one would expect a social response to a VH is similar to that

⋆ Peter Khooshabeh was supported by a fellowship from the Cognitive Sciences Branch of the Army Research Lab Human Research and Engineering Directorate.
⋆⋆ This research was sponsored by the National Science Foundation (grant # IIS-1211064) and the US Army RDECOM. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 339–348, 2013.

elicited by a real human. In the real world, if an individual makes a mistake, others are apt to criticize him or her. How do humans respond to a VH's mistake? In this study we observe how interactants respond to VHs' mistakes in a problem-solving scenario.

In the broader view, our study seeks to inform theories of VH design. The Threshold Model of Social Influence (TMSI) [6,7] is one theoretical model that aims to explain how aspects of a VH's design impact its effectiveness. According to this model, the extent to which people are socially impacted by a VH depends on three variables. *Agency* is the extent to which the participant believes he or she is interacting with another sentient human being. *Communicative realism* (formerly *behavioral realism* [6]) is the degree to which human representations and objects behave as they would in the physical world, which consists of *movement realism*, *anthropometric realism*, and *photographic realism*. Finally, *response system level*, which varies from automatic to deliberative, characterizes the extent to which the interaction with a VH involves shallow automatic cognitions, or deeper deliberate ones. With these three factors, the model argues that the influence of a VH is heavily impacted by its communicative realism, unless interactants believe it is controlled by another sentient human, or if interactants are not thinking deeply about the interaction. Our study, manipulating communicative realism and observing individual differences, serves as a test on the validity of the TMSI.

Communicative realism includes not only the extent to which it looks human (i.e., anthropometric realism and photographic realism) but also how it acts (i.e., movement realism) and speaks [8]. From this perspective, the errors produced by current state-of-the-art natural language processing serve to reduce the realism of the VH's behavior. Previous works studying VHs' social influence have typically focused on errorless VHs [9,10]. In this study we independently manipulate VHs' conversational errors. Based on the TMSI, we predict that erroneous VHs' social influence, if any, will be less than that of errorless ones.

Regarding the *photographic realism* component, the TMSI suggests that visual fidelity should have modest contribution when compared with behavior [6]. Indeed, a later study [9] demonstrated that a non-photorealistic VH can be as persuasive as a photorealistic one. However, there is disagreement on this point: some other studies have found that *photographic realism* enhances V-VH interaction [11]. To look into the debate, the present experiment manipulates *photographic realism* independently.

While communicative realism refers to aspects of the VH, the third factor of the TMSI - response system level - refers to aspects of the human interactant. Response system level could be impacted by the nature of the task in which the human and VH are engaged (e.g., a thought-provoking puzzle vs. an improvised conversation). The factor can also be influenced by the mind-set of the human interactant. For example, some people are inherently deep-thinkers, whereas some others tend to act intuitively and automatically. Such distinction is addressed by social psychologists as the Need For Cognition (NFC) [12]. More recently and consistently with the TMSI, NFC has been proposed as one of the factors explaining when people will anthropomorphize non-human entities [13]. Specifically, people low in NFC are more likely to engage with VHs, leading to the speculation that such people are more influenced by VHs.

The literature on NFC suggests to further unpack the effect of NFC on social influence. It has been reported that individuals low in NFC (i.e., intuitive thinkers) tend to focus on peripheral cues, instead of content of the information [14]. In a H-VH persuasion dialog, presumably the content of the dialog is central, whereas aspects of communicative realism (such as photorealism and behavioral realism) are peripheral, at least to the extent that they do not modify the content of the persuasive dialog. Therefore we can expect an interaction between communicative realism and NFC: when presented with departures from communicative realism, individuals low in NFC are likely to overlook the underlying persuasive information, and become less persuaded.

Other than factors presented by the TMSI, other variables may play an important role as well. For instance, a VH's and interactant's gender have been found to interact during social scenarios. Zanbaka and colleagues reported that people are more affected by VHs of the *opposite* gender [9], whereas Guadagno found an opposite pattern of results [10]. This study manipulates VH gender to revisit the discrepancy.

To summarize, the purpose of this study is to determine:

1. Whether an error-prone VH is capable of social influence during a persuasion task.
2. How do potentially interfering variables, including VH's photographic realism, VH's gender, interactant's NFC, and interactant's gender, interact and moderate VH's social influence.

2 Method

2.1 Participants

Three-hundred and twenty-six workers were recruited from Amazon Mechanical Turk and participated in our study for monetary reward. All workers resided in the United States. There were 48% females and 52% males, with an average age of 31.5 (SD=10.8). Caucasians comprised 71% of the sample, whereas the remainder consisted of Asians (13%), Hispanic Americans (8%), and African Americans (6%). One participant was excluded from analysis because the measured social influence was more than three SD's away from the mean.

2.2 Material

We assessed VHs' social influence by adapting a standard laboratory task that was designed to measure persuasion. In the Lunar Survival Scenario [15,16], the participant is asked to imagine he or she is stranded on the moon and need to prioritize a set of items necessary for survival. The participant can discuss the priorities with a teammate, and possibly make changes as a result of the conversation. Persuasiveness is measured by the extent to which the participant shifts his or her priorities in the direction of the teammate's advice.

The VHs in our study acted as the teammate and provided advice consistent with NASA's expert opinions. The conversation was text-based: the participant asked a question by selecting from a pre-determined set of questions, and the VH responded via printing text on the screen. A screenshot of the interface is shown in Fig. 1. To make the VH appear less mechanic, a delay of 2.5 seconds was introduced before the VH answered each question.

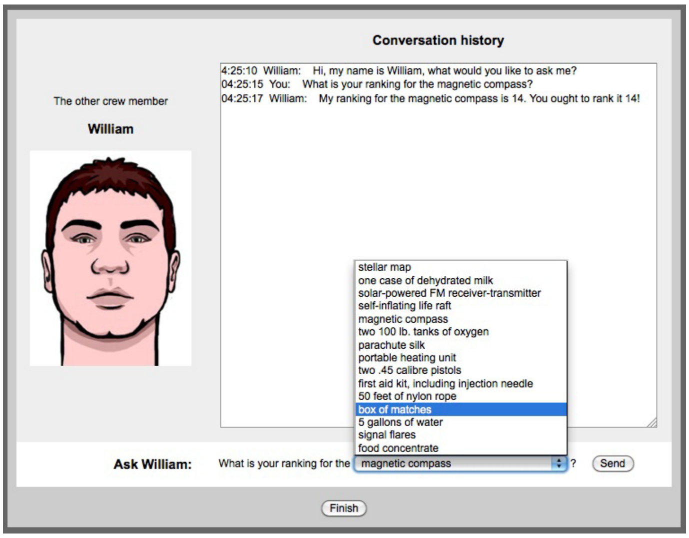

Fig. 1. The H-VH interaction interface

About half the participants were assigned an error-prone VH as the teammate. An error-prone VH makes a conversational mistake 33.33% of the time when he or she answered a question. The type of mistakes is randomized among three patterns [4]:

No response. The VH does not reply to the participant's question. An empty line is printed in place of un expected answer.

Wrong answer. The VH gives information about an irrelevant item, e.g.:
H: What is your ranking for the ***box of matches****?*
VH: I ranked ***oxygen*** *1. You should rank it 1 if you haven't done so!*

Repeat. The VH answers the same question twice, for instance:
H: What is your ranking for the stellar map?
VH: My ranking for the stellar map is 3. Rank it 3!
VH: My ranking for the stellar map is 3. Rank it 3!

Despite the introduction of conversational errors, the information provided by the VH is always accurate, i.e. the VH always gives correct rankings. In situations where the VH is talking about an irrelevant item (such as *oxygen* in the example above), the ranking of that item (i.e., oxygen) is accurately given.

2.3 Design

This study used a between-subject design with three independent variables: Error {Errorless, Erroneous} X Photographic realism {Photorealistic, Non-photorealistic} X VH's Gender {Female, Male}. Social influence is measured via three dependent variables:

Correctness of Final Rankings. How well a participant's *final rankings* match the VH's rankings (i.e., NASA's expert rankings), with 100 being a perfect match, and lower scores indicating further deviation from the VH's rankings, in terms of Euclidean Distance:
$Correctness = 100 - \sqrt{\sum_{i=1}^{15}(H'sRanking_{item_i} - VH'sRanking_{item_i})^2}$

Improvement Score. How much a participant's rankings have shifted toward the VH's rankings, calculated as the subtraction of the Correctness of initial rankings by the Correctness of final rankings:
$ImprovementScore = Correctness_{InitialRankings} - Correctness_{FinalRankings}$

Change Score. How much a participant's rankings have changed, defined as the difference between the initial and final rankings, in terms of Euclidean Distance. (The difference between *Improvement Score* and *Change Score* is that *Improvement Score* is a signed measure, with a positive value indicating the amount of changes made in the direction of the VH's suggestions, whereas *Change Score* is an unsigned variable reflects the amount of total changes, both toward and against the VH's advice.)

$$ChangeScore = \sqrt{\sum_{i=1}^{15}(InitialRanking_{item_i} - FinalRanking_{item_i})^2}$$

2.4 VH Faces

The purpose of this study is to carefully manipulate communicative realism while holding other factors - the ones that could impact persuasiveness - constant. Previous work has emphasized that people judge by facial appearance and act differently [17,18]. To control for appearance, we used an independent set of workers to rate the perceived personality of a number of male and female VH faces. We selected pairs of faces that were perceived to have the same personality. The procedure is described below.

Three male and three female VHs were chosen from an institutional database. For each VH, one photorealistic and one non-photorealistic facial image were created, in a way that outlines of the two faces overlapped. One hundred workers, recruited through Amazon Mechanical Turk, rated each face (in randomized order) on twenty-seven traits. Twenty-four traits are personality traits from [19], whereas the remaining three are: *Honest*, *Familiar*, and *Photorealistic*. Factor analysis on the twenty-seven traits yielded three main factors:

Trustworthy/Honest, which accounts for 33.00% of the variance; *Clever/Wise*, which explains 15.84% variance; and *Introvert/Unsociable*, contributing to 8.25% of the variance. All three factors have a Chronbach's alpha greater than 0.8.

Ratings in terms of the three main factors revealed that, one male VH's photorealistic and non-photorealistic faces appeared to convey the same personality. One female VH's photorealistic and non-photorealistic faces were considered similar: the first two factors achieved consistent ratings, and the third factor resulted in slightly different ratings ($M = 0.07$ vs. $M = -0.10$; $t(195) = 2.1, p = 0.034$; Cohen's $d = 0.3$, effect size= 0.15). These two VHs (shown in Fig. 2) were selected for later study.

A manipulation check on photorealism was performed. The female VH's photorealistic and non-photorealistic representations were rated 3.0 (SD=1.1) and 2.1 (SD=1.1) respectively, on a scale of 1 to 5, with 5 meaning *Very photo-realistic*. The ratings are significantly different with $t(198) = 5.12, p < 0.001$. The male VH's photorealistic version had a mean rating of 3.2 (SD=1.0), which differs dramatically from his non-photorealistic representation ($M = 2.2, SD = 1.2; t(198) = 6.40, p < 0.001$).

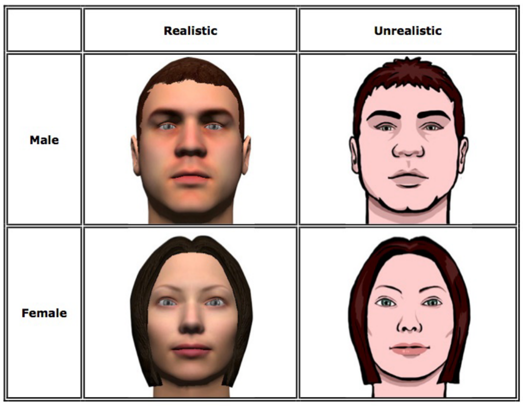

Fig. 2. VHs used in the study

2.5 Procedure

The participants were directed to our online questionnaire, which started with basic demographic questions including age, gender, ethnicity, and the Need for Cognition Scale (NFCS).

Next, participants were informed that they had crash landed on the moon and needed to choose items in order to trek 200km back to a life-saving rendezvous point. Additionally, they were told that another crew member was also present. Due to the crash landing, the captain had been incapacitated and the participant was now the officer in charge. Participants were told that the the other crew member, who knew the ship's inventory well, would be able to answer questions about the item rankings.

Although the other crew member knew the inventory well, the instructions stressed that the participant was the officer in charge. As the captain, the participant had to make the final decision on how to rank the items. We intended the participants to have the decision power and responsibility in the scenario so that they would not blindly take or reject the virtual agent's advice.

Participants initially ranked items prior to chatting with the virtual agent by dragging and dropping the task items on a graphical user interface. After the participants completed their pre-chat rankings, they were reminded that they were in charge and now had an opportunity to chat with the other crew member, who was a VH. Participants had to decide how they would use the information that the VH told them in order to make a set of post-chat rankings.

3 Results

Social influence. Participants' rankings changed significantly after receiving persuasive information from the VHs (within-subjects effect $F(1, 323) = 37.6, p < 0.001$). Participants' rankings shifted toward VH's rankings in both the errorless condition ($F(1, 184) = 113.0, p < 0.001$), and erroneous condition ($F(1, 139) = 37.6, p < 0.001$). These results imply that VHs, even erroneous ones, are capable of social influence.

Error. Error has an effect on *Improvement Score* ($F(1, 323) = 4.4, p = 0.037$) and *Correctness of Final Rankings* ($F(1, 323) = 10.9, p = 0.001$). As predicted, erroneous VHs are less influential (Improvement Score: $M = 2.7, SD = 5.2$ vs. $M = 3.9, SD = 5.0$; Correctness of Final Rankings: $M = 84.9, SD = 5.9$ vs. $M = 87.0, SD = 5.5$).

NFC. We performed a median split on participants' NFCS scores ($Median = 21.0$). The means of the two groups are significantly different ($M = -3.5, SD = 19.0$ vs. $M = 40.0, SD = 12.2$; $t(323) = 24.3, p < 0.001$).

We found an effect of NFC on *Change Score* ($F(1, 323) = 8.1, p = 0.005$), but not on *Improvement Score* or *Correctness of Final Rankings*. Participants low in NFC made more changes ($M = 13.0, SD = 6.7$) than those high in NFC ($M = 11.0, SD = 5.6$).

A three-way interaction between $ErrorXPhotorealismXNFC$ is found (Improvement Score: $F(1, 317) = 5.9, p = 0.010$; Correctness of Final Rankings: $F(1, 317) = 5.9, p = 0.016$). In the context of low-NFC individuals interacting with non-photorealistic VHs, conversational errors have an effect on persuasiveness (Improvement Score: $F(1, 317) = 12.0, p < 0.001$; Correctness of Final Rankings: $F(1, 317) = 12.1, p = 0.001$). Individuals who experienced errors were less persuaded. This observation is consistent with the error effect found on the whole sample. However, this is the only cell that shows an error effect, implying that it is driving the overall error effect.

A second cell in the three-way interaction reveals a NFC effect: under the condition that the VHs are non-photorealistic and errorless, NFC has an effect on persuasiveness (Improvement Score: $F(1, 317) = 12.9, p < 0.001$). Low-NFC individuals are more persuaded than high-NFC individuals under this condition.

Photographic realism. Photographic realism alone does *not* have any effect on VHs' social influence.

Gender. The expected two-way interaction between *VH's Gender* and *Participant's Gender* was *not* found. Instead, we observed an effect of *Participant's Gender* on *Change Score* ($F(1, 323) = 7.6, p = 0.006$), but not on *Improvement Score* or *Correctness of Final Rankings*. In short, female participants made more changes ($M = 13.0, SD = 6.0$) than male participants ($M = 11.1, SD = 6.4$).

4 Discussion

In a problem solving scenario, we measured VHs' social influence by observing interactants' attitude change after receiving persuasive information. To study the effect of design characteristics of VHs, we manipulated communicative realism (via introducing conversational errors), photographic realism, and VH's gender. Our study suggested:
1. VHs who make conversational mistakes are capable of social influence.
2. Individual differences in participants' response system level (indexed by NFC), together with communicative realism (both photographic and linguistic realism), play a role on VH's social influence.

The TMSI suggests that individuals who use deliberative cognitive processing when responding in socially motivated scenarios have relatively higher thresholds of social influence. The observations in this study in the errorless, non-photorealistic condition confirmed this prediction: participants high in NFC (i.e. deep-thinkers) were *less* influenced, whereas users low in NFC were *more* influenced.

For individuals low in NFC (i.e., intuitive thinkers), their decision differed significantly when facing errorless vs. erroneous VHs. As noted above, studies [20,14] have reported that individuals low in NFC tend to focus on peripheral cues. The conversional errors in our study are peripheral to the content of the conversation, suggesting that individuals low in NFC were distracted by errors and overlooked the underlying persuasive information; consequently, they were not persuaded. However, as demonstrated by our data, error alone was not sufficient to produce this effect. Rather, conversational errors combined with non-photorealistic appearance drove the observed effect. These results provide support for the TMSI with regard to low-NFC participants whose response-system level is more automatic.

In comparison, the high NFC group were impervious, both to the manipulation of conversational errors and VH's photorealism. This is possibly due to the fact that individuals high in NFC tend to focus on the content of the information rather than peripheral cues [14]; they may, hypothetically, have filtered out conversational errors and photographic differences from underlying persuasive arguments. As a result, they processed the arguments as if peripheral changes were nonexistent. These observations on high NFC participants align with the TMSI. Together with the findings for low-NFC participants, the results stress the importance of the response system level factor in the TMSI.

Our results may shed some insight into the mixed findings concerning photorealistic vs. non-photorealistic VHs. Observations from this study is consistent with the interpretation that visual fidelity may be insufficient in and of itself to alter the effectiveness of VHs. Rather, it may interact with other factors in determining social influence. In our experiment, conversational errors combined with non-photorealistic appearance demonstrated an effect on individuals low in NFC. One possible explanation is that the error plus non-photorealism combination produced a significant effect that would have been less apparent if the factors were to be introduced in isolation. Alternatively, the impact of photorealism might have overridden the effect of errors for individuals low in NFC. Either way, our findings emphasize the importance of separately controlling for different aspects of communicative realism in future studies.

Another characteristic of the interactants we investigated was gender. The expected interaction between VH's gender and participant's gender was absent in our study. More research is in demand to justify whether people are more affected by VHs of the *opposite* gender [9], or the *same* gender [10].

There are several future directions that we are interested in pursuing. In the present study, we controlled for communicative realism and response system level, but not *agency*. It is possible that conversational errors not only impacted realism, but also participants' willingness to attribute human sentience to the agent. To determine the impact of errors on agency, we need to independently manipulate *agency*.

The concept of NFC also calls for further exploration. In this work we analyzed between-subject differences in NFC in a repeated task. Additionally, we can manipulate within-subject variations in NFC by having the same subject performing different tasks that vary in cognitive needs.

While the present study examined factors from one theoretical model regarding VH influence (i.e., TMSI), other models exist in the literature. For example, the Ethopoeia concept of Nass and colleagues [21] has been raised as an alternative model for guiding VH design. Although our results lend support to the TMSI, we did not directly contrast TMSI with other models. More research is in demand to guide the theoretical foundations of VH design.

References

1. Bickmore, T., Gruber, A., Picard, R.: Establishing the computer–patient working alliance in automated health behavior change interventions. Patient Education and Counseling 59(1), 21–30 (2005)
2. Woolf, B.P.: Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning. Morgan Kaufmann (2010)
3. Rizzo, A., Buckwalter, J.G., Forbell, E., Reist, C., Difede, J., Rothbaum, B.O., Lange, B., Koenig, S., Talbot, T.: Virtual reality applications to address the wounds of war. Psychiatric Annals 43(3), 123–138 (2013)
4. Skarbez, R., Kotranza, A., Brooks, F., Lok, B., Whitton, M.C.: An initial exploration of conversational errors as a novel method for evaluating virtual human experiences. In: 2011 IEEE Virtual Reality Conference (VR), pp. 243–244. IEEE (2011)

5. Nass, C., Steuer, J., Tauber, E.R.: Computers are social actors. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 72–78. ACM (1994)
6. Blascovich, J.: Social influence within immersive virtual environments. In: The Social Life of Avatars, pp. 127–145. Springer (2002)
7. Blascovich, J., McCall, C.: Social influence in virtual. In: The Oxford Handbook of Media Psychology, p. 305 (2013)
8. Khooshabeh, P., Dehghani, M., Nazarian, A., Gratch, J.: The cultural influence model: When accented natural language spoken by virtual characters matters. Journal of Artificial Intelligence and Society
9. Zanbaka, C., Goolkasian, P., Hodges, L.: Can a virtual cat persuade you?: the role of gender and realism in speaker persuasiveness. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1153–1162. ACM (2006)
10. Guadagno, R.E., Blascovich, J., Bailenson, J.N., McCall, C.: Virtual humans and persuasion: The effects of agency and behavioral realism. Media Psychology 10(1), 1–22 (2007)
11. Groom, V., Nass, C., Chen, T., Nielsen, A., Scarborough, J.K., Robles, E.: Evaluating the effects of behavioral realism in embodied agents. International Journal of Human-Computer Studies 67(10), 842–849 (2009)
12. Cacioppo, J.T., Petty, R.E.: The need for cognition. Journal of Personality and Social Psychology 42(1), 116–131 (1982)
13. Epley, N., Waytz, A., Cacioppo, J.T.: On seeing human: a three-factor theory of anthropomorphism. Psychological Review 114(4), 864 (2007)
14. Haugtvedt, C., Petty, R.E., Cacioppo, J.T., Steidley, T.: Personality and ad effectiveness: Exploring the utility of need for cognition. Advances in Consumer Research 15(1), 209–212 (1988)
15. Johnsen, K., Raij, A., Stevens, A., Lind, D.S., Lok, B.: The validity of a virtual human experience for interpersonal skills education. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1049–1058. ACM (2007)
16. Khooshabeh, P., McCall, C., Gandhe, S., Gratch, J., Blascovich, J.: Does it matter if a computer jokes. In: PART 1-Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems, pp. 77–86. ACM (2011)
17. Wilson, R.K., Eckel, C.C.: Judging a book by its cover: Beauty and expectations in the trust game. Political Research Quarterly 59(2), 189–202 (2006)
18. Stirrat, M., Perrett, D.I.: Valid facial cues to cooperation and trust male facial width and trustworthiness. Psychological Science 21(3), 349–354 (2010)
19. Hassin, R., Trope, Y., et al.: Facing faces: Studies on the cognitive aspects of physiognomy. Journal of Personality and Social Psychology 78(5), 837–852 (2000)
20. Petty, R.E., Cacioppo, J.T.: The elaboration likelihood model of persuasion. In: Communication and Persuasion, pp. 1–24. Springer (1986)
21. Nass, C., Moon, Y.: Machines and mindlessness: Social responses to computers. Journal of Social Issues 56(1), 81–103 (2000)