The Faces of Health: Exploring the Relationship Between Anthropomorphic Tendencies and Perceived Source Credibility in Digital Avatars

Kyle Qian
Stanford, USA
kyleqian@stanford.edu

Will Kim
Stanford, USA
wkim@cs.stanford.edu

Lucio Tan
Stanford, USA
lctan21@stanford.edu

Kevin Yang
Stanford, USA
kyang14@stanford.edu

ABSTRACT

With the recent rise of artificially-intelligent companions (e.g. Siri, Facebook M, and Alexa), engineers and designers are putting a face to new technology. Cognizant of this future, we believe it is crucial that we understand the effect of avatars on people and their perception of credibility in artificially-intelligent applications. In this paper, we explore the relationships between specific anthropomorphic themes (humanoid, companion, inanimate object) and an avatar's perceived source credibility (factors include competence, caring/goodwill, and trustworthiness). We recruited 57 subjects through Mechanical Turk to interact in a controlled setting with a web-based health assistant, which takes the form of one of six randomly assigned avatars. Post-interaction surveys reveal that subjects assigned to humanoid avatars perceived the avatar to be less intelligent, less informed, and more self-centered than did subjects assigned to other avatars.

ACM Classification Keywords

Human Factors; Design; Experimentation

Author Keywords

Anthropomorphism; Credibility; Bots; Digital avatars

INTRODUCTION

R2D2, Wall-E, and Baymax are beloved fictional depictions of artificial intelligence for a reason. Once a face is added to a technology, all of a sudden it becomes more human, more credible. With recent years, this dream is becoming more and more real. Artificially-intelligent companions, like Siri, Facebook M, and Alexa, have been introduced to the world as early attempts of this advanced form of human-computer interaction. Hence, as designers and engineers, it is important that we understand how different faces of avatars affect people's psychology.

In the context of artificially-intelligent applications, avatars are the interface between the application and its end users. Thus, they hold great influence on the end users' perception towards an application, its features, and its brand. With this in mind, we wanted to explore the relationship of avatars and the perceived credibility of artificially-intelligent applications. In the spectrum of appearances, what are the common avatar categories? How do they each affect the components of credibility and credibility as a whole?

Our study attempts to discover any potential mapping, dependency, or correlation between an avatar's anthropomorphic tendency scale category and the resulting source credibility factors.

RELATED WORK

Our research is inspired by prior work on the correlations between the appearance of virtual avatars and its effect on user perception. Garau Slater's "The Impact of Avatar Realism and Eye Gaze Control on Perceived Quality of Communication in a Shared Immersive Virtual Environment" (2003) explored the effects of facial realism and eye gaze behavior on perceived effectiveness of communication. Their paper concluded that for realistic avatars, inferred eye gaze behavior was more effective than random eye gaze behavior; for unrealistic avatars, inferred eye gaze behavior was actually less effective than random eye gaze behavior. In Nowak Rauh's "The Influence of the Avatar on Online Perceptions of Anthropomorphism, Androgyny, Credibility, Homophily and Attraction", participants were asked to evaluate a set of avatar faces in terms of characteristics such as androgyny, anthropomorphism, credibility, homophily, and attraction. It was discovered that anthropomorphic avatars were perceived to be more attractive and credible, that feminine avatars were more attractive than masculine avatars, and that most subjects preferred human avatars that matched their gender.

Our research primarily builds on the work of Nowak Rauh, in that our experiment involves varying avatar appearance to

CS376, December 14, 2016, Stanford, CA, USA

ISBN 978-1-4503-2138-9.

DOI: 10.1145/1235

influence the way it's perceived by a subject. But as opposed to having subjects evaluate our avatars in a static context, we set up an interactive session between the subject and avatar, a method used by Garau Slater.

Chin et al has developed a 208-item scale to better understand humans' anthropomorphic perceptions of objects in "Developing an Anthropomorphic Tendencies Scale." This research finds that there are four general themes of anthropomorphic tendencies: 1) extreme anthropomorphic tendencies, 2) anthropomorphic tendencies towards a spiritual being, 3) anthropomorphic tendencies towards pets, and 4) "negative" anthropomorphism towards non-human entities. From this paper, we inform the design of our six avatars, ensuring complete coverage of these four categories ("human", "sentient being", "pet" ,and "inanimate objects"). Effectively, this paper serves as the foundational framework for the independent variables of our experiment when taken in conjunction with McCroskey's "Source Credibility Measures."

J. C. McCroskey offers a structured approach to thinking about source credibility by breaking it down into component factors. His approach integrates a multidimensional measure of three final factors: competence, caring/goodwill, and trustworthiness. Each of these three factors are themselves determined by the cumulative score of their constituent atomic factors (e.g. intelligence, honesty, expertise in topic, etc.), which are rated on a Likert scale of one to seven. We apply findings from this research when creating our survey, which ask participants to rate their avatar on each of these atomic factors.

As stated above, the works of Garau Slater and Nowak Rauh provided us with foundational background knowledge of current research concepts and methodologies involving the correlation of avatars and the end user experience. We extend upon these concepts by leveraging the more quantitative work of Chin et al and McCroskey to provide concrete metrics for our our avatar designs and credibility measures.

EXPERIMENT DESIGN

Overview

Our work leverages prior research in anthropomorphism, source credibility, and virtual avatar design to investigate A.I. chatbot interfaces in the context of health. Our final experimental system/procedure first involved interaction with a simple web chat system, where users converse with our health assistant and receive one recommendation towards a healthier life, followed by a likert-scale based credibility survey. We will now provide specific design details of our experimental system and insights that led us to make these decisions.

Avatar Design

We based the appearances of our six avatars (spiritual being, man, robot, dog, apple, and circle) on the work of Chin et al. By anchoring our designs on this robust scale, we explore a wide spectrum of avatar faces and how they affect credibility when interacting with a user. Further, these six bot avatars were then grouped into three major categories: robot and dog were companion avatars, circle and apple were inanimate avatars, and person and spiritual being were humanoid avatars.

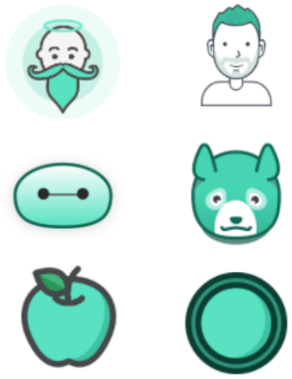


Figure 1. An example of our six bots, without the blinking/hovering animation. Each row is a cohort, from top down: humanoid, companion, inanimate

Furthermore, we created the avatars with modern web design guidelines in mind. All avatars have a consistent, 2D visual appearance across color, complexity, scale, gender, and vector style.

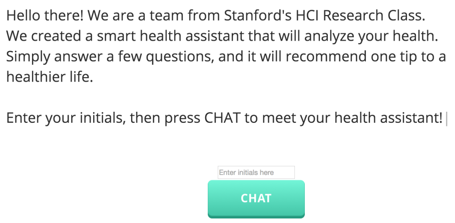
Beyond the face, animation is crucial for users to feel like they are speaking to a responsive avatar assistant. Early on after pre-pilot tests, we learned the experience felt closer to filling out a form than an actual conversation. We were often told that "the bot felt unresponsive" or that they "thought the picture was just [our] logo." Consequently, several people failed to remember the avatar at all. In order to humanize our health assistant conversation, we first introduce it to the user. For instance, our intro page ends with "Press CHAT to meet your health assistant," and we animated the avatar to blink its eyes and move up and down in a bobbing movement. These measures increased users' identification and awareness of our avatars.

Test Procedure

Subjects are brought to the study through an online link introducing our avatar as a "smart health assistant." After submitting their initials to be saved as a session token, the subjects arrive at a page featuring their randomly assigned avatar and a simple chat box. The subjects would then proceed to converse with the avatar.

Following a script that's held constant across all avatars, the health assistant would (1) introduce itself, (2) ask basic patient health information, (3) ask six general health questions, and (4) based on the subject's answers, provide a generic health recommendation.

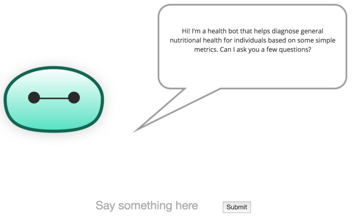
After the subject receives their health recommendation, they are linked directly to a Qualtrics survey. The survey is designed to measure perceived source credibility as a combination of the three factors detailed by the "Source Credibility Measures" paper: competence, caring/goodwill, and trustworthiness. In addition to basic demographic information, subjects filled out an 18-point Likert scale questionnaire (six questions per factor) to evaluate their interactions with the avatar. After



Hello there! We are a team from Stanford's HCI Research Class. We created a smart health assistant that will analyze your health. Simply answer a few questions, and it will recommend one tip to a healthier life.

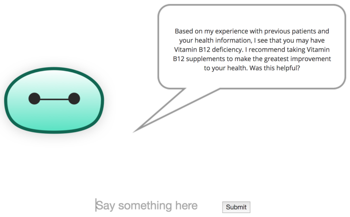
Enter your initials, then press CHAT to meet your health assistant!

Figure 2. Intro to the bot. This is done in an animated form as if someone is writing this out.



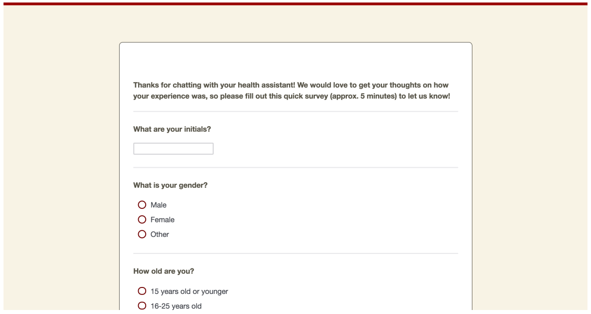
Hi! I'm a health bot that helps diagnose general nutritional health for individuals based on some simple metrics. Can I ask you a few questions?

Figure 3. The actual bot chatting



Based on my experience with previous patients and your health information, I see that you may have Vitamin B12 deficiency. I recommend taking Vitamin B12 supplements to make the greatest improvement to your health. Was this helpful?

Figure 4. The bot's recommendation after the full chat (not shown)



Thanks for chatting with your health assistant! We would love to get your thoughts on how your experience was, so please fill out this quick survey (approx. 5 minutes) to let us know!

What are your initials?

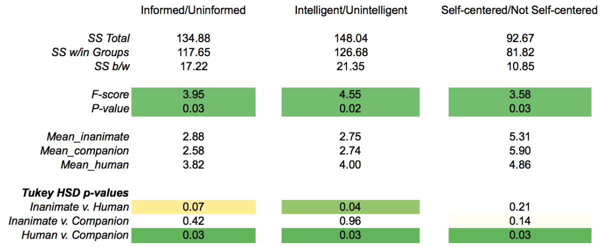
What is your gender?

○ Male
○ Female
○ Other

How old are you?

○ 15 years old or younger
○ 16-25 years old

Figure 5. The post-chat survey on Qualtrics



	Informed/Uninformed	Intelligent/Unintelligent	Self-centered/Not Self-centered
SS Total	134.88	148.04	92.67
SS w/in Groups	117.65	126.68	81.82
SS b/w	17.22	21.35	10.85
F-score	3.95	4.55	3.58
P-value	0.03	0.02	0.03
Mean_inanimate	2.88	2.75	5.31
Mean_companion	2.58	2.74	5.90
Mean_human	3.82	4.00	4.86
Tukey HSD p-values			
Inanimate v. Human	0.07	0.04	0.21
Inanimate v. Companion	0.42	0.96	0.14
Human v. Companion	0.03	0.03	0.03

Figure 6. Statistically significant results for each question, with post hoc Tukey analysis p-values also included. Color-coded green if statistically significant with p<0.05. Df (numerator): 2; Df (denominator): 54

the survey, subjects were shown a debrief of our study, specifically noting that our generic health recommendation is not to be taken seriously.

Participants

We recruited a total of 57 participants (roughly 10 participants per avatar) through Mechanical Turk to interact with our health assistant avatar and to take a post-interaction survey. Participants were paid $1 per completed interaction and survey, a rate of $6-12 per hour depending on how long they took.

EXPERIMENTAL RESULTS

A one-way ANOVA was conducted to compare the relative effects of each anthropomorphic theme on subjects' perception of source credibility, measured as a combination of competence, caring/goodwill, and trustworthiness. These three factors are derived from the 18-question survey that users took after interacting with the bot, where three groups of six questions correspond to each factor of the source credibility. These initial results showed a significant difference between the three bot categories with regard to the competence factor of source credibility [F(2, 54) = 3.32, p = 0.04]. Although the one-way ANOVA yielded a statistically significant result, post hoc comparisons using the Tukey Honest Significant Difference (HSD) test did not yield significant differences between any pairs of anthropomorphic themes with regard to competence.

In order to distill the core drivers behind the high variance between the three groups, we then conducted one-way ANOVA tests on the specific questions that respondents filled out (18 total per survey), grouping by the bot type (humanoids, companions, inanimate objects). We found that there were significant differences between the bots on users' perceptions of whether the bot was intelligent, self-centered, and informed.

We found that the human bot faces were seen as less intelligent than either companion bots or inanimate bots ($p<0.05$). Further, we found that the human bots were seen as more self-centered and less informed than the companion bots ($p<0.05$). Otherwise, for questions regarding competence, sensitivity, and the like, we have failed to show that there is any statistically significant difference between any pair of the bot types.

DISCUSSION

Our early results showed significant three-way differences between our avatar categories with regard to the competence factor of source credibility. Upon drilling down to the level of individual questions, we found that this three-way difference was mostly driven by users' perceptions of whether the bot was intelligent, informed, and self-centered, the first two of which are components of the competence factor of source credibility. Specifically, we found that the humanoid avatars were perceived as less intelligent than both the companions and the inanimate objects. The humanoid avatars were also perceived as less informed and more self-centered than the companion avatars.

Thus, given rather simplistic tasks and conversational structure (form-like chat), we've demonstrated that a humanoid bot will perform worse than its companion bot and inanimate bot counterparts in demonstrating that it is indeed well-informed. In these settings, a humanoid bot will also perform worse than the companion bot when compared on intelligence and selfishness. This may be due to the stilted nature of the conversation, compared to the myriad human interactions that the user may be used to (while the user may feel that the conversation with the companion bot was more novel).

While we cannot say that choosing a non-humanoid avatar will necessarily result in greater perceived credibility, we can conclude that given two avatars (one humanoid and the other a companion) with the same script, it is likely that the non-humanoid (yet anthropomorphized) bot will score higher on intelligence, non-self-centeredness, and competence.

LIMITATIONS

It was difficult to make the avatar noticeable within a web interface, hence the users' expressed perceptions did not excite nearly as strong an arousal as we'd have liked. However, this could change for future applications with avatars (i.e. virtual reality agents where the avatar is a direct person we are interacting with or for robots).

Further, using Mechanical Turk workers who are incentivized by time led to several results being immediately tossed (i.e. we had one "test" question that asked subjects to select the fifth option), but we kept others that were of questionable value because it was not clear whether the Mechanical Turker was truly "undecided" across the board or gaming the survey. Finally, we structured our Likert scale such that the middle point (4 in the scale of 1 to 7) was also said to denote "undecided." As a result, it is possible that the users may have felt more often than not that they could not judge a humanoid bot as well (because they were perhaps treating it more like a human than other bots, who then require more attention and time to make judgments). To mitigate this for future studies, it may be worthwhile to make an explicit choice for "uncertain," and/or not denote the middle point as "uncertain."

FUTURE WORK

Currently, we've displayed the perceptions of humans in the physical world interacting with robots in a virtual environment. However, with future technologies, the use cases will expand and as virtual reality agents/robots become more widely used, it becomes ever more important to consider which types of agents elicit stronger emotions of perceived trustworthiness, competence, and self-centeredness. We imagine that there are many more emotional responses that beings/things will elicit and are excited to see the results of more study in this space.

CONCLUSION

In this study, we ran one-way ANOVA tests on the results of 57 users chatting with three different types of bots (companion, humanoid, inanimate). We found that these different types of bots do elicit different perceptions on credibility (composed of caring, trust, and competence scores), but fail to have any pairwise significance (both with $p<0.05$). We drilled down further to see the main drivers behind the differences in perceptions, finding that there were statistically significant differences between the perceived intelligence and self-centeredness of humanoid bots (generic man and spiritual man) and companion bots (dog and Baymax–a cute robot). Surprisingly, users found that humanoid bots were less intelligent and more self-centered than companion bots. Further, we also found that there were statistically significant differences whether the bot was informed or not, where users judged the humanoid bots to be less informed than both the companion and inanimate bots ($p<0.05$). We believe this may be due to the higher expectations for human competency combined with the stilted nature of the scripted conversation (which we kept exactly the same in order to reduce the number of confounds). We think this research may be valuable for those moving forward in developing virtual reality and mobile health applications, and are excited to see future work done to demonstrate the applicability of these learnings to more intelligent, conversational experiences.

ACKNOWLEDGMENTS

We thank Michael Bernstein, Rob Semmens, and Kesler Tanner for their active help in guiding this paper to its final form. Hope we're not killing Rob this time around.

REFERENCES

1. Chin, Matt G., et al. Developing an Anthropomorphic Tendency Scale (2005): Proceedings of the Human Factors and Ergonomics Society (49): http://journals.sagepub.com/doi/pdf/10.1177/154193120504901311

2. GARAU, M., et. al (2003). The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. In CHI. 2003: ACM Press.

3. K.L. Nowak, C. Rauh (2005). The influence of the avatar on online perceptions of anthropomorphism, androgyny, credibility, homophily, and attraction. Journal of Computer-Mediated Communication, 11 (1), pp. 153âĂŞ178. http://dx.doi.org/10.1111/j.1083-6101.2006.tb00308.x

4. McCroskey and Teven (1999). A reexamination of the construct and its measurement. Communication Monographs. Source Credibility Measures. Online: http://www.jamescmccroskey.com/measures/source _credibility.htm

5. Walters, M.L., Syrdal, D.S., Dautenhahn, K. et al. (2008). Avoiding the Uncanny Valley: Robot Appearance, Personality and Consistency of Behavior in an Attention-Seeking

Home Scenario for a Robot Companion. In Auton Robot 24: 159. doi:10.1007/s10514-007-9058-3