Virtual Humans Elicit Skin-Tone Bias Consistent with Real-World Skin-Tone Biases

Brent Rossen[1], Kyle Johnsen[1], Adeline Deladisma[2], Scott Lind[2], Benjamin Lok[1]

[1]CISE
University of Florida
Gainesville, FL 32611, USA
{brossen, kjohnsen, lok}@cise.ufl.edu

[2]Dept of Surgery Oncology
Medical College of Georgia
Augusta, GA 30912, USA
{adeladisma, dlind}@mail.mcg.edu

Abstract. In this paper, we present results from a study that shows that a dark skin-tone VH agent elicits user behavior consistent with real world skin-tone biases. Results from a study with medical students (*n*=21), show participant empathy towards a dark skin-tone VH patient was predicted by their measured bias towards African-Americans. Real world bias was measured using a validated psychological instrument called the implicit association test (IAT). Scores on the IAT were significantly correlated to coders' ratings of participant empathy. This result indicates that VHs elicit realistic responses and could become an important component in cultural diversity training.

Keywords: Virtual humans, intelligent agents, virtual reality, human-centered computing, user interfaces, racial bias, medicine, computer graphics.

1 Introduction

Virtual human (VH) agents have been shown to be useful in many fields such as training military skills [1], social conversation protocols [2], and clinical therapist skills [3]. These interactions with VHs focus on interpersonal goals that are fundamentally similar to human-human (H-H) interactions. It is known that real world biases (e.g. race, age, gender, weight) impact real world interactions [4]. Thus we seek to ask the same questions about virtual world interactions. First, do biases affect VH interactions, and second, are these VH biases correlated with real world human (H) biases. Specifically, we study the effect of VH skin-tone on the cognition and behavior of a human conversational partner. The expectation is that by changing only the VH's skin-tone, the user's bias towards or against a particular race will cause a modification in the user's behavior.

We explore the behavioral effect of VH skin-tone in the context of the Interpersonal Simulator (IPS). The IPS allows users to interact naturally (speech and gestures) with life-sized VHs [5]. Currently applied to interpersonal skills training, the IPS has been used by over 450 health professions students to practice medical interviews. The eventual goal of the IPS is to train students on areas that are difficult with current techniques, e.g. cultural diversity training. Using the health-care domain as a test-bed empowers the study of diversity related issues. Medical students are motivated to improve, and the interview is a stressful experience (important for eliciting biases). Further, the health-care domain is one where skin-tone biases are an established problem [6]. Conventional techniques for interview training (e.g. role playing, actors) have difficulty providing the diversity and repeated exposures necessary for cultural diversity training. Many of these logistical hurdles in cultural diversity training would be addressed through the integration of VH agents.

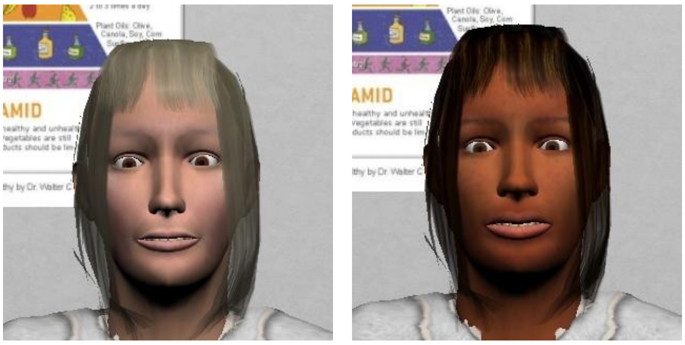

Fig. 1. The two VH skin-tones. The average skin-tone on the left is "Light" <201, 152, 138>, and on the right is "Dark" <112, 58, 32>.

Before integrating VHs into cultural diversity training, we must establish their efficacy for cultural diversity education. To establish this efficacy, we need to link measured real world biases to cognitive and behavioral effects in H-VH interactions. Cultural bias metrics come in two forms, direct and indirect. Direct measures assume users can accurately recognize their own biases. Surveys (e.g. "Does a person's race determine how much you like them?") are often used for this purpose. A known issue with direct measures of bias is that of social desirability bias. Social desirability bias is observed when people answer questionnaires in a way that accommodates social expectations, rather than how they truly perceive a situation. To circumvent this issue, indirect measures are used (i.e. measures where the user does not know that bias is being measured, or cannot control their bias). Indirect metrics include *behavioral observation*, *indirect survey questions* ("How empathetic were you?"), and *implicit response latency measures*. One implicit response latency measure, the Implicit Association Test (IAT) is widely used by psychologists for establishing skin-tone biases. The IAT has been validated in the United States general population for race, age, gender, and weight with very large data sets (N = 40,000 to 160,000 each) [7]. Using *direct* measures to study bias, one can never be sure if the measure has resulted in the desired bias or has retrieved only the *socially desirable* answer. The current study evaluates bias using *indirect measures* to evaluate the underlying racial biases.

The present work conducted a study (n=21) comparing medical students' interactions with a light-skin (Average <Red, Green, Blue>= <201, 152, 138>) and a dark skin-tone (<112, 58, 32>) VH patient (See Fig. 1). The objectives were to determine the applicability of bias measures, and identify potential effects of bias in the medical interview training test-bed. This study is a first step towards validating VH agents for cultural diversity education.

2 Previous Work

Efforts have been made to establish the basis of effective VHs and validate their efficacy in social interaction. This study adds to the growing body of literature on highly interactive and emotionally engaging VHs, and cultural diversity training applications [1, 2]. It continues the research by examining the effect of one of the more subtle elements of VHs, their skin-tone.

Skin-tone has been shown as a factor in the persuasive power (a cognitive marker of bias) of a VH. Studies have shown that users are more persuaded by VHs that emulate the user's own race [8]. They have also shown that learning from a VH is significantly affected by agent race [9]. The present work builds upon these results by studying cognitive *and* behavioral markers of skin-tone bias in a bidirectional *conversation* with a VH.

3 Study Design

The study was a between-subjects experimental design with VH skin-tone as the sole independent factor. Two conditions were tested, light skin-tone VH and dark skin-tone VH. Medical students performed a typical patient interview with either a light skin-tone VH *or* a dark skin-tone VH.

Medical students are taught to follow a structure when conducting a patient interview. The student's direct goal is to obtain a history of the illness and form an initial differential diagnosis for the patient. The protocol of a patient interview is explored in under ten minutes while building rapport with the patient by expressing empathy (understanding and concern for the patient's problems).

One VH patient, Edna, was used in the study. Edna represented a 55 year-old female who has found a lump in her left breast. Edna's voice was pre-recorded audio of an actor (a 75-year old Caucasian woman with a local accent) and was used for both skin-tone appearances.

Edna's rendering and interaction were accomplished using 3 networked computers. A wireless microphone captured student speech and allowed un-tethered natural language interaction. Optical tracking allowed appropriate gaze behavior. The VHs had realistic body meshes and employed skeletal and morph animations. Considerable efforts were made to mimic the patient-doctor interaction. The students interacted with life-size VH patients in a clinical skills examination lab. No icons, desktops, mice, or keyboard were ever visible. More details about the IPS hardware, infrastructure, and believability can be found in [5].

3.2 Population

Three Caucasian first-year physician assistant (PA) students and eighteen Caucasian third-year medical students attending a medical college in the south-east United States participated in this study. Group assignment was pseudo-random to ensure that each group had a similar distribution. The dark skin-tone condition had *n*=12 with 3 females and 9 males, the light-skin condition had *n*=9, with 4 females and 5 males.

Originally, recruitment did not take participant race into consideration, but only small numbers of non-Caucasians were in each condition. As a result, *only Caucasian participants were used for analysis* (the reason for the condition size disparity) because they were the largest homogenous group available. Examining just that group provides more information on bias, particularly out-group bias (bias against anyone not within the same group as the participant), than achieving a higher number of participants (original *n*=27, current *n*=21). This is a technique used in psychology studies such as [4].

3.3 Metrics

The Implicit Association Test (IAT): The IAT is a system for determining subconscious bias [7]. It measures response latencies, requiring the participant to rapidly categorize stimuli into "good" and "bad" categories. Detailed information and demonstrations can be found at [4, 7] and the Project Implicit website. It has been shown that survey metrics can be controlled by participants *motivated* to avoid biased responses, while this motivation has no effect on the IAT [4]. The result of the IAT is the IAT-D value. The ***IAT-D*** value is a number in the range -1 to +1, wherein .15=slight, .35=moderate, and .65-strong preference. For this study, positive IAT-D values represent a subconscious preference for African-Americans, and negative values represent a preference for Caucasian Americans. A study on racial prejudice by Dovidio et al, uses the IAT and provides excellent details and full literature review of real world bias[4].

Behavioral observation and coding: The medical interview task is a ten-minute interview that largely consists of the participant asking questions and listening to the response from the VH. To facilitate the study of bias a VH initiated challenge was used to encourage a natural, spontaneous response. Participants' resulting behavior was analyzed, as well as participants' post-interview perception of that behavior.

At 4 minutes into the interview the VH, Edna, asked, "Could this be cancer?" this is labeled as the *critical moment*. This critical moment was created to elicit biases based on conversations with medical educators and racial bias researchers (see acknowledgements). Students are expected to respond with *empathy.* Edna's "sister" had cancer and students must show Edna that they understand her concern. Video coders qualitatively rated this moment as "empathetic" on a 7-point scale from 1(not at all) to 7 (very). Observers reviewed the critical moment using 1) Audio only, 2) Video only, and 3) Audio + Video. Separating audio and video created a division of participants' verbal (e.g. speaking words of understanding) and non-verbal behavior (e.g. leaning towards the patient, or making a facial expression). Reviewers were blind to the race of the VH. Inter-rater agreement is reported in the results section.

Post-experience survey: Following the interview, participants were asked to rate various aspects of their whole interview (overall performance and use of empathy). Racial bias may play a role because the ratings indicate how much a participant enjoyed the interview process. Participants were unaware that skin-tone bias was being tested, thereby limiting the effects of social desirability bias.

3.4 Procedure

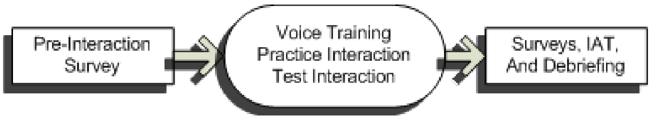

Fig. 2. H-VH Procedure overview

Upon arrival, the participant completed a consent form and the background questionnaire. The background questionnaire was used to collect information about the previous experience, interview skill, gender, race, and age of the participant.

Following a 2-minute voice training for the speech recognition system, the participant was asked to walk into the isolated section of the clinical skills lab where the virtual experience was set up, and have a seat on the chair. Each participant performed one VH patient interview. A short introduction to spoken interaction with a VH was then given by one experimenter. The participant then performed the patient interview alone. Afterwards, the participant was escorted back to the survey room. There they took the post-experience survey and were then given the IAT. Finally, they were debriefed about the purposes of the experiment. Each session took one hour.

4 Results

Given a population with a **real world bias** similar to the US national average, the hypothesis is that participants in the dark skin-tone condition will demonstrate behavior indicative of bias against the VH, and participants in the light skin-tone condition should demonstrate behavior indicative of bias favoring the VH. **Empathy** should be higher in the light skin-tone condition and lower in the dark skin-tone condition. Furthermore, we expect a positive correlation of real world bias (the IAT) to empathy in the dark skin-tone condition, and a negative correlation of real world bias to empathy in the light skin-tone condition.

Real World Bias: The study population had a slight to moderate racial bias against African-Americans (M=-0.26, SD =.41) according to the IAT results. Seventeen (81%) of the participants had a negative score, and five had a positive score. IAT scores for the participants in the dark skin-tone condition (M=-.33, SD =.35) were not significantly different than IAT scores for the light skin-tone condition (M= -0.17, SD=.46). This distribution is similar to national averages (M=-0.30, SD=0.83) [7].

Empathy: Coders (2 audio-only, 1 video-only, 2 audio and video) rated clips of user behavior after the VH challenged the participant with the question "Could this be cancer?" An inter-rater reliability analysis was conducted, and found that rater reliability was acceptable (average intra-class correlation coefficient = 0.894).

As predicted, real world bias was positively correlated to self-reported and observed empathy in the dark skin-tone condition, with significant (p<.01) correlations to Audio-only (r=0.84), Audio-video (r=0.74), and Average (r=0.79) observed empathy. In the light skin-tone condition, the results were not as clear.

Unexpectedly, real world bias was not significantly correlated (p>.10) to self-reported (r=-.01) or observed empathy (avg r=0.18) in the light skin-tone condition.

Also, significant (p<.01) *negative* correlations were observed between self-reported and observed empathy (r=-.91) in the light skin-tone condition. This is an unexpected result given the significant (p<.05) *positive* correlations (r=.60) in the dark skin-tone condition. In the light skin-tone condition, the more students used empathy (as rated by observers), the lower they rated themselves. This result suggests that racial bias impacts the internal scale participants use to rate themselves.

As expected, empathy was higher in the light skin-tone condition. A significant (Wilk's λ=0.481, F(3,16)=5.75, p<.001) multivariate effect of skin-tone condition was found for the degree of empathy observed during the empathetic moment. The average of all coders showed that the dark skin-tone patient was empathized with less (M=2.98, SD=1.52) than the light skin-tone patient (M=4.05, SD=1.67). Fig. 3 shows the average empathy scores given by each group of coders for the dark skin-tone and light skin-tone groups.

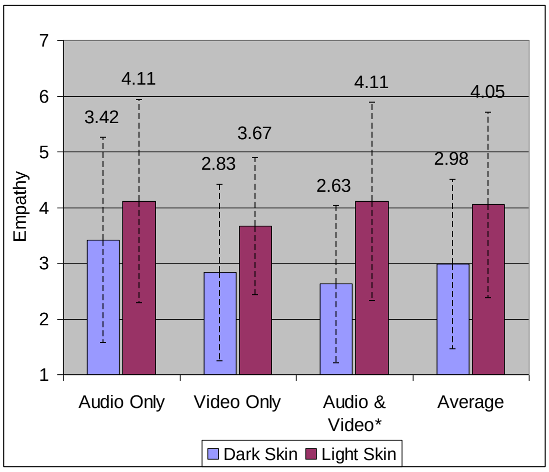

Fig. 3 Average scores for observed empathy during the empathetic moment. Empathy was rated on a scale from 1 (not at all) to 7 (very). (* p<.05)

5 Discussion

The results of this study indicate that VH agent skin-tone is a factor influencing both behavior and self-perception of empathy during patient interviews. Trends were observed which showed more empathy being used by participants in the light skin-tone condition than by participants in the dark skin-tone condition. Furthermore, real world bias (IAT score) is significantly correlated with empathetic behavior for participants in the dark skin-tone condition, which indicates that the two metrics share underlying factors. Establishing this correlation is the first step towards proving the validity of behavioral bias metrics with VHs.

For participants in the light skin-tone condition, real world bias (the IAT) did not appear to be a predictor of empathetic behavior, as little correlation was observed. This correlation was expected to be significantly negative. Instead, for many coders we observed a positive correlation. Further investigations as to why this occurred are ongoing.

6 Conclusions and Future Work

In this paper, we reported results from a study, which examined the capability of VH agents to elicit behavior consistent with real world skin-tone bias. We found that participants expressed more empathy towards a light skin-tone VH than a dark skin-tone VH. This correlated to the study population's real world bias favoring Caucasian-Americans over African-Americans as measured by the IAT.

This bias translated to a task in the form of empathetic behavior. During a medical interview of a VH patient, the patient challenged the participant with "Could this be cancer?". The skin-tone of the patient affected participant behavior while responding to this challenge. From the results of the IAT, we expected to find more empathy exhibited with the light skin-tone VH patient than the dark skin-tone patient. Indeed, participant's self-reported empathy and coders' ratings of empathy confirmed this expectation. Furthermore, the real world bias significantly predicted participants' empathy in the dark skin-tone condition.

In conclusion, we find that real world H-H biases transfer to the virtual world in H-VH conversations. We have demonstrated that behavior with VHs is correlated to real world racial bias using a validated measure, the IAT. These results indicate that VHs elicit racial bias in conversational tasks, and motivate the use of VHs in experiences where it is desirable to elicit bias as an educational tool, e.g. in military, medicine, and business cultural training. Educators in these fields aim to identify and mitigate bias, and VH exposures can become a powerful tool to augment current training techniques.

Next, this study will be repeated with a larger number of participants, at least 30 in each group. It will be a within-subjects study where each participant interacts with both a dark skin-tone and a light skin-tone VH. We will also examine if the trends exhibited with this Caucasian group hold true with other ethnic groups.

Assuming the larger N study of skin-tone bias confirms what was indicated in this study, we will examine if extensive exposure to dark skin-tone VHs has a mitigating effect on racial bias. Last, it will be explored whether other biases can be elicited from the appearance of a VH. For instance, it would be useful to know if the apparent age, gender, and weight of the VH have a predictable effect on user behavior.

Acknowledgments

Special thanks go to Dr. John Dovidio and Dr. Ashby Plant for their advice on eliciting racial bias as well as their reviews and support of this work. We also thank Andrew Raij and Aaron Kotranza for their assistance conducting the user study. This work was supported by University of Florida Alumni Fellowships and a National Science Foundation Grant IIS-0643557.

References

1. Deaton, E., Barba, C., Santarelli, T., Rosenzweig, L., Souders, V., McCollum, C., Seip, J., Knerr, W., and Singer, J., “Virtual Environment Cultural Training for Operational Readiness (Vector).” *Virtual Reality,* vol. 8, no. 3, pp. 156-167, 2005.
2. Babu, S., Suma, E., Barnes, T., and Hodges, L. F., "Using Immersive Virtual Humans for Training in Social Conversational Protocols in a South Indian Culture."
3. Kenny, P., Parsons, T., Gratch, J., Leuski, A., and Rizzo, A., "Virtual Patients for Clinical Therapist Skills Training." pp. 197-210.
4. Dovidio, J. F., Kawakami, K., and Gaertner, S. L., “Implicit and Explicit Prejudice and Interracial Interaction,” *Journal of Personality and Social Psychology,* vol. 82, no. 1, pp. 62-68, 2002.
5. Johnsen, K., Dickerson, R., Raij, A., Lok, B., Jackson, J., Shin, M., Hernandez, J., Stevens, A., and Lind, D. S., "Experiences in Using Immersive Virtual Characters to Educate Medical Communication Skills." pp. 179-186, 324.
6. Cohen, J. J., Gabriel, B. A., and Terrel, C., "The Case For Diversity In the Health Care Workforce. Project HOPE."
7. Nosek, B., Mahzarin, B., and Greenwald, A., “Harvesting Implicit Group Attitudes and Beliefs from a Demonstration Web Site,” *Group Dynamics: Theory, Research, and Practice,* vol. 6, no. 1, pp. 101-115, 2002.
8. Pratt, A. J., Hauser, K., Ugray, Z., and Patterson, O., “Looking at Human-Computer Interface Design: Effects of Ethnicity in Computer agents,” *Interacting with Computers,* vol. 19, no. 4, pp. 512-523, 2007.
9. Baylor, A. L., and Kim, Y., “Pedagogical Agent Design: The Impact of Agent Realism, Gender, Ethnicity, and Instructional Role,” *Proc. of International Conference on Intelligent Tutoring Systems*, 2004.