

# The Impact of Avatar Stylization on Trust

Ryan Canales\*  
Reality Labs Research, Meta  
Clemson University

Doug Roble  
Reality Labs Research, Meta

Michael Neff  
Reality Labs Research, Meta  
University of California,  
Davis



Figure 1: Images from our VR study showing the three identities (AF, AM, CM) and styles (Real, Mid, Caricature).

## ABSTRACT

Virtual Reality (VR) affords great freedom in how one represents themselves in virtual interactions through the selection of different avatars. However, it remains unclear which avatar should be chosen for a given social scenario. Social interaction often relies on the establishment of trust. Are people more likely to trust you if you select a highly realistic avatar or is there flexibility in representation? This work presents a study exploring this question using a high stakes medical scenario. Participants meet three different doctors with three different style levels: realistic, caricatured, and an in-between “Mid” level. Trust ratings were largely consistent across the style levels, but participants were more likely to select doctors with the “Mid” level of stylization for a second opinion. There is a clear preference against one of the three doctor identities, with evidence that this may be related to movement features.

**Index Terms:** Human-centered computing—Human computer interaction (HCI)—HCI design and evaluation methods—User studies; Computing methodologies—Computer graphics—Graphics systems and interfaces—Virtual reality

## 1 INTRODUCTION

It is becoming clear that embodied representations are important for effective interaction in VR (e.g. [34, 35]), and appearance can vary broadly, from cartoon to realistic to nonhuman. The impact of appearance on social interactions is not well understood, but despite conflicting evidence, there is a tendency to view realism as important [41]. Realism imposes high costs in terms of the required technology and may not always be aesthetically preferable. In this paper, we present a study in which we investigate the influence that avatar stylization has on trust. Trust is a key ingredient in social interaction, and an attribute for which realism is postulated as being particularly important.

\*e-mail: rcanale@clemson.edu

Our study employs the trust model proposed by Mayer et al. [24] which posits that trust requires people to be vulnerable and that trusting another relies on positive perceptions of their ability, benevolence and integrity. In our study, vulnerability is introduced by having participants experience a high stakes scenario in which they must imagine that they have received a cancer diagnosis and must select a doctor for a second opinion. They interview three doctors using a VR system, with pre-recorded responses tailored to address the three components of the trust model we employed. Each doctor is a different person (varied *identity*) and each is represented at a different level of stylization: realistic, caricature, or a mid condition intended to be in-between realistic and caricature. While providing variation in identity is important for maintaining ecological validity, and scenario plausibility for a within subjects design, our primary research focus is on the impact of stylization. In particular, we focus on professionally crafted caricatures that vary both the sculpt and texture of the character, but maintain the look of the actual person.

In order to ensure behavioral consistency across participants and to limit the technical complexity, we used prerecorded avatar animation rather than live tracking. This allowed us to have very high quality motion, using well established offline motion capture techniques for hands, body, and face. It also ensured consistency across the avatars as their full dialogue was set and balanced ahead of time. Participants conducted the interview by selecting from pre-set questions, but could not make open ended queries. Participants completed subjective surveys related to trust and its components after each interview. After interviewing three doctors, they selected one for a second opinion.

Results show very little impact of stylization. Style differences did not lead to any significant differences in the subjective trust scores, and notably, there was no evidence that the realistic avatars were viewed as more trustworthy. Participants were more likely to select the Mid level of stylization when picking a doctor for the second opinion. While this stylization did not significantly outperform the other stylizations in subjective survey results, it is postulated that Mid may have eliminated some issues with the more extreme stylizations that could have lead to subconscious discomfort or allowed better alignment between the actor motion and avatar.

There was a strong impact of identity, with one doctor being rated lower on trust surveys and also selected less frequently for the second opinion.

## 2 BACKGROUND

### 2.1 Avatar Stylization

We use “Stylization” as a general term for the variation in the look of an entire class of avatars, rather than choices made for a specific avatar, such as hair color or clothing. Style classes include photorealistic, cartoon-like, semi-realistic, etc. and include shape, texture, and rendering variations. They may refer to well known types of work, like anime, or aesthetic decisions, like the use of rounded, angular or highly caricatured features. There is a growing body of work on the impacts of avatar stylization on user experience, but it is still not well understood, with numerous conflicting findings.

Rating studies show people avatars, sometimes with limited motion, and ask them to rate them on various scales. Inkpen et al. [17] found an inverse correlation between avatars rated as creepy and people’s interest in using them. McDonnell et al. [25] gathered ratings of ten levels of avatar stylization created by changing rendering settings. The most abstract (“toon pencil”) and most realistic were rated highest on appeal and trustworthiness. Fleming et al. [12] blended realistic scanned avatars and hand made stylizations based on movie styles (Pixar, Avengers, etc.). They found for both grey and textured models that the middle blends of 33% or 66% were more appealing than either the realistic or highly stylized. Focusing on faces, Zell et al. [43] combined shape and rendering changes. They found that shape contributes more strongly to perceived realism and material is the main factor in perceived appeal, although moderate shape levels were more appealing in their second study. Realistic materials were least appealing, which may be an uncanny effect. In a meta study, Weidner et al. [41] acknowledge conflicting findings, but argue for realistic avatars. Results may depend on the exact quality of the avatar, for example, for abstract [8] and fairly realistic [9] faces, Ferstl et al. found that narrow faces were seen as more aggressive, but for highly realistic faces, they found the opposite [10], matching the perception of photographs. Other work showed that cultural preferences influence avatar preference, although this can be moderated with role-appropriate clothing [1]. A meta study of agents for eHealth suggests there is a dearth of research on agent appearance [36].

Acting observation studies involve participants rating an avatar while they perform more complex behavior, such as telling a story. There is little evidence that rendering style impacts personality perception. Zibrek and McDonnell [46] compared a realistic, but ill character to a cartoon shaded character and found the cartoon shaded character was seen as having a more agreeable personality, but ratings were not otherwise significantly different. Ruhland et al. [33] found that personality can be conveyed from head and eye movement on both realistic and cartoon models. Zibrek et al. [44] compared the perception of personality across a range of rendering styles. There were some effects of rendering, but no clear pattern emerged. Lighting was shown to alter the intensity of emotions and brighter lighting approved appeal [42]. In an agent storytelling scenario, Zibrek et al. [45] found greater empathetic concern for a Sketch rendering of an avatar within a sad scenario, but Concern was highest for a realistic style in a friendly scenario. Other work showed that photorealism did not improve empathy, but did improve social presence, although it remained quite low [47]. McDonnell et al. [25] performed a lie detection task with a range of rendering conditions as well as real video and audio-only and found no significant differences in lie detection between conditions. Ring et al. [32] found that cartoon proportions are seen as more friendly, but realistic proportions are seen as more appropriate for medical applications. Ferstl et al. [11] found that for likability, a character with real voice, full motion capture, and a robot embodiment was rated highest, exceeding a

realistic embodiment. For anthropomorphism ratings, there was no difference between the robot and realistic character, despite one being a robot.

Studies on advice giving avatars that compared video and hand animated avatars found that only the advice text, and not the avatar appearance, changed people’s opinions, even though the animated avatar was rated as more eerie [29]. Similar work with a longer medical scenario found that the animated doctor actually increased adherence intention (small effect) and consultation enjoyment [4]. In follow up work with more visual conditions, it was found that agent behavior, rather than appearance, influenced outcomes [5]. From a study of different renderings of a sick medical patient, researchers conclude that a less visually realistic appearance may actually elicit a stronger emotional reaction related to the scenario, but visual realism is still important for eliciting social-emotional constructs such as shyness and shame [39].

Behavior can lead to differences in user performance. Torre et al. [37] found that people trusted an agent more that smiled (vs. neutral expression) with a neutral voice (vs. happy). In a follow up study that added a cartoon agent, only small differences in performance were found, but the photorealistic avatar was rated more realistic and more eerie, the cartoon more appealing, attractive, and happier. Patoskaya et al. [30] found people changed their navigation behavior to avoid a neurotic character more than an emotionally stable one.

A number of studies have focused on manipulating people’s personal avatar. A visually matched avatar [40] or realistic avatar [20] can improve factors like body ownership, although not in all studies [21]. People will also moderate their physical behavior with more realistic avatars [14, 28]. Immersion and enjoyment can increase when the avatar is aligned with the story [6]. Ma and Pan [22] compared realistic and cartoon avatars, matched to the participant, and found appearance was rated higher for the cartoon model, but body ownership was higher for whichever avatar they used first. No effect was found on gambling behavior for avatar clothing designed to evoke socioeconomic differences [27]. Jo et al. [18] found co-presence was higher with a cartoon character than a realistic one on a job interview task, but that a realistic avatar was seen as more trustworthy on a knowledge based task. On a story telling task, Radiah et al. [31] found that personalized avatars can lead to higher body ownership and agency, as well as emotion elicitation. Customizing a person’s own avatar (racially matched hands), led to an increase in pro-social behavior [15], but watching a speech by an avatar matched to the user increased perceptions of uncanniness, leading to a decrease in trust.

Motion has a strong impact on the perception of avatars, as evidenced by studies that vary motion quality and through people’s prioritization of motion control when customizing their own avatars [13]. There is also evidence that the motion signal should not be degraded [6, 19] even when other signals, such as voice or visual appearance, are degraded [11]. Normoyle et al. found that gaze behavior influences the perceived trustworthiness of avatars, with averted gaze being viewed as less trustworthy than direct eye-contact [26].

### 2.2 Trust

Trust is seen as a key ingredient for effective communication and collaboration. Hossain and Wigand argue that “the real challenge for the management of virtual collaboration is trust” [16]. If trust is established, people will share information openly, but without it, transactions must be monitored closely to avoid exploitation [3]. Without trust, workers may adjust tasks to avoid the need for close collaboration or avoid collaborating altogether. Trust supports more efficient work and adapting to changing circumstances [3].

The concept of trust is multifaceted. Cognitive-based trust relies on the rational judgment of a person’s knowledge, competence, and dependability whereas affect-based trust relates to an emotional bond

and the belief that another is protective of your interests and shows genuine care for your welfare. For a recent survey of trust evaluation, please refer to [38]. For this work, we rely on the model of trust proposed by Mayer et al. [24] that defines trust as “the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party” [p.712]. The definition posits that in order for a person to have a need to trust, they must be vulnerable, so trust involves the willingness to take a risk. They divide trust into three factors. *Ability* relates to a person’s skills and competence such that they can complete the desired task. *Benevolence* relates to a belief that the trustee will act in the best interest of the trustor. *Integrity* requires that the person acts according to a set of principles that the trustor finds acceptable. Trust is also modulated by an individual’s propensity to trust.

### 3 SCENARIO

We wished people to have a plausible interaction with avatars in order to explore their trust formation across different avatar representations. The trust literature suggests that in order for people to truly experience trust, they must feel a sense of vulnerability or there is no need to trust [24]. We thus needed a scenario that made participants feel vulnerable. After pre-testing several high stakes scenarios (Section 3.1), we adopted a medical scenario that was rated as being both relatable and causing vulnerability. In the selected scenario, participants were told that a lump had been found in their kidney that might be cancerous. Before they sought treatment, they wanted to get a second opinion. In this experiment, they interviewed three different doctors and selected one to provide that additional opinion. The interviews took place in what they were told was a new VR-based telecommunication system. In this system, a doctor would appear and users could ask him/her three different questions by clicking on the button with the associated text. After the third question, a tube would rise up around the doctor to “transport” them away and the next doctor would appear.

#### 3.1 Scenario Testing

A set of potential scenarios was brainstormed by the research team and workshoped with colleagues. Based on this process, five scenarios were selected for evaluation by users through an online survey using Mechanical Turk. These scenarios included having to defend against embezzlement from a small business, a medical scenario in which you feel unwell, a medical scenario involving a lump in your kidney, the need for advice investing large lottery winnings, and a medical scenario involving a cancer diagnosis of your child. MTurk workers reviewed paragraph long descriptions of each scenario (included in Sec. 12 of the Appendix) and then rated their agreement of the following prompts: “I can imagine what it would be like to be in this scenario.” and “Imagining myself in this scenario makes me feel vulnerable.” on 7-point Likert scales.

The survey was completed by 39 workers (27 male, 12 female; mean age  $M=42.8$ ,  $SD=10.5$ ). Our goal was to find a highly relatable scenario, so all of our future participants could imagine themselves in the situation, and one that also made people feel vulnerable. We considered means, statistical tests of differences, and also inspection of the histograms. The kidney scenario has the highest mean ( $6.6/7$ ,  $SD=0.97$ ) for vulnerability, with the embezzlement ( $M=6.33$ ,  $SD=0.93$ ) and a child with cancer scenario ( $M=6.4$ ,  $SD=1.2$ ) performing similarly. The lottery winnings investment scenario ( $M=4.95$ ,  $SD=1.8$ ) and feeling unwell ( $M=5.36$ ,  $SD=1.4$ ) performed significantly worse than the other three and were eliminated from consideration. The kidney scenario had the highest relatability of the retained scenarios ( $M=5.8$ ,  $SD=1.1$ , compared to childhood cancer  $M=5.2$ ,  $SD=1.7$ , embezzlement  $M=5.4$ ,  $SD=1.4$ ). There was a tendency for childhood cancer to be rated lower ( $p=.077$ ,

Tukey method; higher than  $p=.05$ ). Inspecting the histograms, the kidney scenario only had three ratings below the neutral value of 4 and they were all at the “slightly disagree (3)” level. The other two options also had three ratings at this level, but in addition had five ratings at “disagree (2)” for childhood cancer and two ratings at this level for the embezzlement scenario. This suggests that we may be more likely to run into people that have difficulty relating to these two scenarios. We selected the lump in the kidney scenario as it had good relatability and engendered a strong sense of vulnerability.

#### 3.2 Dialogue Testing

Each interview with a doctor includes responses to three questions that map to the three constituents of trust in our model: ability, benevolence, and integrity. For this baseline study, we wanted all three doctors to perform at a high level on these three components of trust. If a particular avatar stylization outperforms others, it would be interesting to conduct follow up work to intentionally manipulate the trustworthiness of the dialogue to see how this interacts with the impact of stylization.

In order to maximize realism and ensure authenticity, the avatar’s voice must match the avatar, which means that a different actor must perform each avatar’s dialogue. During the scenario, each avatar must also provide different answers to ensure believability. We must ensure that each avatar has equivalent dialogue that conveys similar levels of trust. To achieve this, we pretested a large set of written answers to each of the ability, benevolence, and integrity questions in online surveys. We then selected dialogues that were statistically equivalent. Final assignments were done to also ensure a match with the avatars (e.g. a dialogue that mentioned 25 years of experience was assigned to the oldest doctor, AM).

For ability, we wrote 17 different responses and had 40 participants rate them on Mechanical Turk (16 female, 24 male; mean age 38.3,  $SD=9.0$ ). Participants were given the background scenario about the lump in the kidney and then for each dialogue sample rated their agreement to two prompts “This person has the ability to the job;” and “I can imagine a doctor responding like this” on 7-point Likert scales. The first prompt measures the perceived ability and the second is a safety check to ensure that the dialogue will be believable. A weak response was included as a manipulation check, but most text was designed to portray a high but believable level of ability.

The lowest mean rating for ability was 2.7 and highest 6.5 on the 1 to 7 scale. Believability means ranged from 2.1 to 6.2. Three dialogues were selected that had ratings that were not statistically distinguishable and were in the second to highest rating group out of six groups: Dialog 12 ( $M=6.3$ ,  $SD=0.78$ ; assigned to AF), Dialog 9 ( $M=6.1$ ,  $SD=0.97$ ; assigned to AM) and Dialog 7 ( $M=6.1$ ,  $SD=1.3$ ; assigned to CM). These are included in the appendix (Sec. 13). They were also in the second highest group for plausibility with means 5.8, 5.8 and 5.6 respectively.

For benevolence, we wrote 15 different dialogue samples and had them rated by 35 participants (13 female, 22 male; mean age 40.3,  $SD=9.2$ ). The content prompt was changed to “This person will behave in a benevolent (well meaning) manner.” with the same believability prompt.

Ratings for benevolence ranged from a mean of 2.6 to a mean of 6.2 and believability from 3.1 to 5.9. The chosen dialogues were in the second highest rated group of seven and not statistically distinguishable: Dialog 5 (mean 5.9,  $SD=1.1$ ; assigned to AF), Dialog 3 ( $M=5.7$ ,  $SD=1.0$ ; assigned to AM) and Dialog 14 ( $M=5.9$ ,  $SD=1.2$ ; assigned to CM). Their believability ratings fell in the second highest group and were not statistically distinguishable, with means 5.3, 5.7 and 5.6 respectively.

For integrity, we wrote 12 sample dialogues and had them rated by 36 participants (16 female, 20 male; mean age 40.6,  $SD=10.2$ ). The content question was “The person adheres to a set of principles

that I find acceptable.” with again the same believability prompt.

The integrity ratings ranged from a mean of 2.0 to a mean of 6.5 with believability ratings ranging from a mean of 2.4 to 6.2. The three selected dialogues fell in the top group of seven based on their Integrity ratings and were not statistically distinguishable: Dialog 9 (M=6.5, SD=0.61; assigned to AF), Dialog 10 (M=6.3, SD=0.86; assigned to AM) and Dialog 12 (M=6.1, SD=0.90; assigned to CM). Their believability ratings fell in the highest group and were not statistically separable, with means 6.1, 6.2 and 5.8 respectively.

## 4 APPARATUS

The main challenge in the study was creating convincing simulations of the doctors at different levels of fidelity. This involved three large efforts: building the avatar models, capturing the performances that would drive the avatars and building an interactive VR experience.

### 4.1 Avatars

We began by selecting three participants to serve as models for the avatars. They varied in age, gender, and racial background, and included an east Asian female (AF) in her late twenties to early thirties, a Caucasian male (CM) in his forties, and a south Asian male (AM) in his fifties. Each actor was recorded in a light stage (e.g. [7]) to capture their head and facial details. The data captured consisted of high-resolution 3D geometric data, color-corrected albedo maps, and normal maps. The facial geometry was then passed through *Mesh to MetaHuman* in order to produce a fully rigged MetaHuman model that could be animated in the Unreal engine. This output model was very close to the input mesh, but contained some small variations due to the lower resolution mesh used by MetaHumans. For example, the details around the eyelids and mouth for the Caucasian model were smoothed out. The bodies for each model were selected from the MetaHuman default body types. We used the medium male normal weight body type for both male avatars and the tall female normal weight body for the female avatar. Likewise, the grooms (hair, eyebrows, and eyelashes) were selected from the MetaHuman set of grooms. We applied our scanned albedo maps to the MetaHuman skin material.

The two stylized versions of each identity were hand modeled by a team of artists. The scanned high realism models and photos of each person were provided to the art team. From this, they created concept art for a caricatured version of each person along with a semi-realistic model intended to lie between the caricature and realistic model. The concept art went through rounds of review with an art director and the art team before converging on the goal models. Three dimensional models were then sculpted based on this 2D concept art. These models also went through several rounds of review. The final models were then passed through the same Mesh to MetaHuman pipeline to produce rigged MetaHuman models. There were limitations to using the Mesh to MetaHuman pipeline with the stylized models, however. Some features, such as the larger eyes and sharper facial features of the stylized models, were not captured by Mesh to MetaHuman, so we had to further manually process the resulting models. We made the Mesh to MetaHuman output mesh fit the shape of the stylized sculpts more precisely using Wrap3D, and also performed additional fine tuning in Maya, such as sculpting and skinning, to reduce artifacts when animating the stylized face. Some preprocessing of the hand sculpted models was done in Maya to make the mesh suitable for wrapping in Wrap3D. We then imported the original high resolution mesh for the head of each model and the Mesh to MetaHuman output into Wrap3D and hand selected corresponding points on each model where we wanted the detail to transfer, such as the sharp angles around the eyes and mouths of the characters. Finally, the wrapped mesh, which is the same topology as MetaHuman, was further tuned in Maya by sculpting the mesh and adjusting the facial joint influences on the geometry. The art team also provided albedo, normal, and roughness maps for

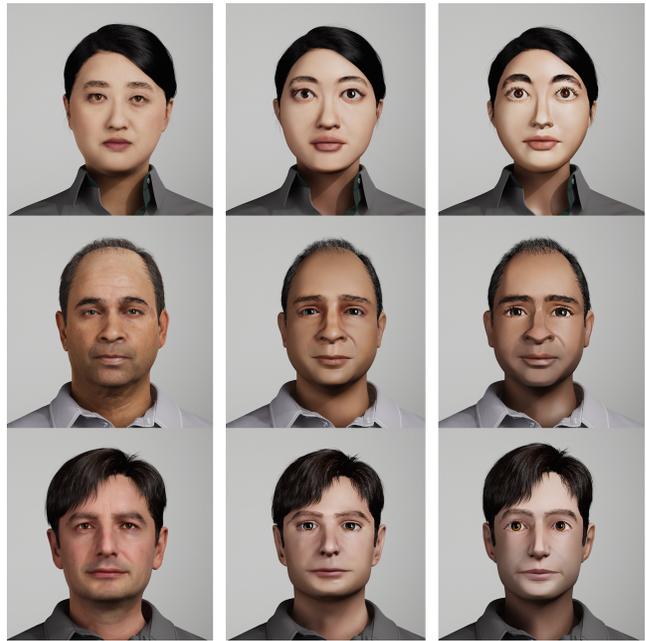


Figure 2: Avatars used in the study. The rows show the three identities, AF, AM and CM. The columns show the three levels of stylization, Real, Mid and caricature.

each stylized avatar, which were applied in Unreal using the default material with tuned parameters. Hair and clothing were added using MetaHuman assets. We aimed to achieve stylized hair while still using the MetaHuman grooms. To this end, we adjusted the size of the individual strands and slightly adjusted where the groom sits on the head, under artistic direction. We used hair cards for the eyelashes of the most stylized versions of the avatars in addition to manual shading of the eyelids to make the eyelashes appear thicker. Clothing was kept neutral and simple across the avatars, consisting of a grey shirt/blouse and black pants. The resulting models rendered in Unreal are shown in Figure 2.

### 4.2 Performance

High quality motion capture was used to animate each doctor. The first step in the recording process was to hire an actor for each of the three identities, with the plan to use the same performance across all levels of stylization for each doctor. Video based auditions were conducted with candidates that matched the race and approximate age of each source model. Candidates read three brief scripts, one was a job interview for a financial advisor position, one bore witness in a criminal case and the third was an appeal to a parole board. Actors were cast based on three criteria: providing high quality readings, having a reasonable match with the facial shape of the target model in order to maximize the chance of successful motion transfer, and with a goal of having performances balanced across the three doctors.

The capture of all actors took place over two days. Performances were directed by a member of the research team with numerous repetitions of each line to ensure quality and correctness. The research team later selected the highest quality takes. A Vicon optical motion capture system was used to record body motion. Actors wore a helmet that held cameras on a bar in front of their face to record facial motion. Post-processing of the face was done using the software developed by Faceware Technologies. The result of this process was keyed animations for the MetaHuman facial control rig and keyed joint angles for the MetaHuman body joints. The output

of this process was the joint angles at each frame for a skeleton with 342 body joints (including correctives) and keyed facial controls at each frame to specify the facial animation.

### 4.3 Experience

The immersive VR experience was created using the Unreal 5 game engine developed by Epic Games. To avoid any potential mismatches between the models and the general setting, a simple grey background was chosen, with some nuance added by light placement. The experience was displayed on a Meta Quest 2 tethered to a PC. All interactions were performed using a visibly rendered virtual ray, originating from either the right or left Meta Quest touch controller (depending on the participant's handedness) and aligned with the depth axis of the controller.

## 5 METHODS

### 5.1 Design

To investigate the effects that avatar style might have on trust, the full experiment uses a 3 (avatar identity) x 3 (avatar style) experimental design, but each participant sees only three avatars, featuring one of each identity and each style (e.g. AM-Caricature, AF-Real and CM-Mid). There are six such combinations that were sampled between subjects. Every possible ordering was shown twice, except for one set that was shown once.

### 5.2 Participants

In total, 67 participants completed the study with one being discarded because of technical issues. The remaining 66 participants ranged from 21 to 65 with a mean age of 34.4 (SD=12.1). Gender was essentially balanced, with 31 identifying as female, 32 male, 1 non-binary/third gender and 2 preferring not to disclose. In terms of race, 33 were White/Caucasian, 11 were Asian/Asian American, 8 were Latin/Hispanic, 5 were Black/African/African American, 1 was Latin/Hispanic, American Indian/Alaska Native, 5 selected combinations of these categories and 3 selected "Other". In terms of VR experience, 21 participants had none, 32 participants had some (1-3 times), 10 used it several times (4+) and 3 owned their own hardware. In terms of education, 9 had graduate degrees, 38 had four year college degrees, 7 had two year college degrees, 6 had some college, 4 had high school diplomas, 1 had some high school and 1 preferred not to disclose.

### 5.3 Procedure

All participants attended a single session. After completing consent, they received the backstory explaining that they were participating in a scenario where they had been diagnosed with a serious kidney issue that was potentially cancerous and they needed to select a doctor in order to get a second opinion. They would be using a novel "telehealth" system, in which the interviewed doctors would be represented as virtual avatars. Participants then put on the Meta Quest 2 headset, and calibrated the virtual space with the help of the research team.

At the start of the VR experience, participants were given instructions and a basic orientation task in order to ensure that they were comfortable interacting in the virtual scene. The orientation task was to select one of three 3D virtual buttons displayed in the scene in front of them by pointing with the virtual ray originating from their controller and using the trigger on the controller to select the specified button. After successfully selecting the correct button, the process was repeated two more times, each with a different button to select. Afterwards, participants were asked to complete a survey on their propensity to trust [23] (See Supplemental Material, Sec. 11.2).

For all surveys, the text was displayed on a white virtual placard in the environment. Participants selected their response by selecting one of seven labeled 3D buttons displayed below the placard, corresponding to the 7-point Likert scale. After a selection was made, a

labeled "Next" button appeared, at which point participants could either change their selection or move on to the next survey question.

After completing the propensity survey, participants were given instructions on interviewing the doctors, then they proceeded to the interview stage of the experience. A white cylindrical barrier ascended from the ground, then a silhouette of the first doctor faded into view briefly, before the barrier retracted back into the floor, revealing the first doctor. Participants could ask the doctor three questions related to their ability, benevolence, and integrity by selecting a virtual placard with the corresponding text. The options were:

- Please tell me about your relevant skills and background.
- How do you balance the needs of your patients with your interests?
- Please describe the set of principles that you adhere to and a time in which those principles were tested.

When the participant clicked on a placard, the pre-recorded audio and animations for the current doctor were played in order for the doctor to deliver an answer. After the doctor finished delivering their response, a customized idle motion was played. Transitions into and out of the animations were achieved with a combination of hand animated transition motions and motion blending. To ensure that eye contact was made when intended by the performer playing the doctor, the avatar's gaze direction was adjusted with additive blending so that the head and upper body would aim at the participant based on the location of their head mounted display.

After all three responses were played, the cylindrical barrier ascended again, this time removing the doctor from the scene. Participants then completed a survey on their impression of the doctor. Prompts were customized from [23] and designed to measure Ability, Benevolence and Integrity, the three constituents of the trust model used in this study, and then a composite measure of Trust. The categories of the prompts were not labeled. The prompts in our study are listed in the Supplementary Material (Sec. 11.1).

After the survey, the next doctor appeared, and the process was repeated twice more so that all three doctors were interviewed and rated. Participants then selected a single doctor for a second opinion, and also selected the doctor they would least prefer. Participants removed the headset and completed a structured interview with the research assistant. They were asked an ordered set of questions that probed their opinion of the doctors and the experience (Sec. 7), with potential follow up questions to clarify comments as needed. Participants were debriefed on the goals of the study after the interviews.

### 5.4 Statistical Analysis

Statistical analysis for all survey data was performed by fitting a linear mixed-effects model to the data using lme4 [2]. Models were fit for the Style factor, Identity factor and Style+Identity with interactions, and the model that best explained the data was adopted. Post-hoc analysis was performed by doing pairwise comparisons with Estimated Marginal Means (emmeans in R). For certain data, other tests were used and will be described in the text.

## 6 RESULTS

People's innate propensity to trust can impact how likely they are to trust in practice. We checked to ensure that this did not skew results. All participants completed a trust propensity survey [23]. There was no significant difference of the propensity of people to trust across any of the nine assigned avatars ( $p=.52$ ). There was also no significant correlation between people's propensity to trust and the average of their actual trust survey ratings as described in Sec.6.1 (Pearson's  $R$  of 0.062,  $p=.62$ ), so it appears that personal trust propensity was not an important factor in the trust results.

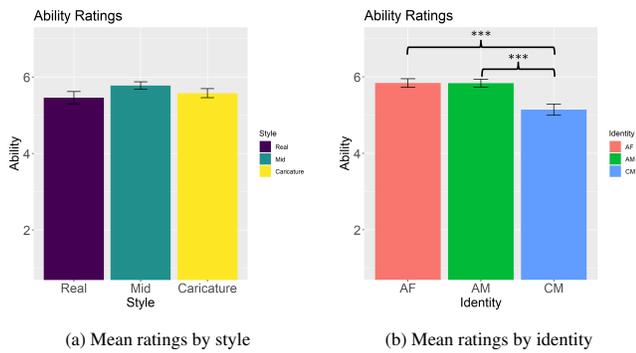


Figure 3: Ability ratings, grouped by Style (a) and Identity (b). Lines point to significantly different pairs. \*\*\* indicates a  $p$ -value less than 0.001.

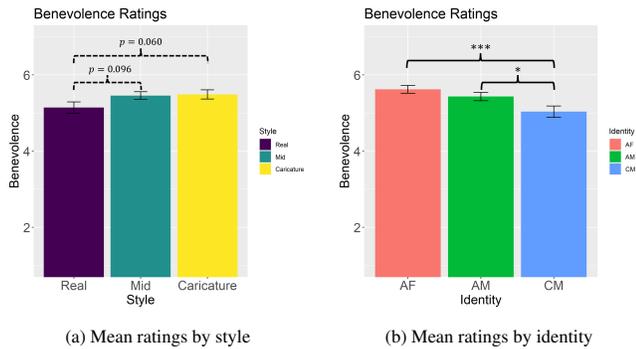


Figure 4: Benevolence ratings, grouped by Style (a) and Identity (b). Lines point to significantly different pairs. \*\*\* indicates a  $p$ -value less than 0.001, and \* a  $p$ -value < 0.05.

## 6.1 Survey Results

Style generally had no impact on ratings, with one partial exception for benevolence. The overall results suggest that the CM identity was rated lower on the various trust measures than the other identities in most cases.

The ability ratings (Fig. 3) show a significant difference for Identity ( $\chi^2(2) = 26.07, p < .0001$ ), but no main effect for Style ( $\chi^2(2) = 4.19, p = .12$ ) and no interaction ( $\chi^2(4) = .82, p = .94$ ). Post-hoc analysis shows that CM had significantly lower ability ratings than AF ( $p = .0001$ ) and AM ( $p = .0001$ ).

Benevolence ratings (Fig. 4) are the only category that show a significant main effect for Style ( $\chi^2(2) = 6.78, p = .034$ ) as well as Identity ( $\chi^2(2) = 16.83, p = .00022$ ), but again, no interaction ( $\chi^2(4) = 5.67, p = .23$ ). Post-hoc analysis shows that CM was rated significantly lower than AF ( $p = .0004$ ) and AM ( $p = .023$ ). Post-hoc analysis did not find significant style differences, but a tendency for Real to be rated lower than Mid ( $p = .096$ ) and Caricature ( $p = .060$ ).

Integrity ratings (Fig. 5) show a main effect for Identity ( $\chi^2(2) = 15.58, p = .0004$ ), but not for Style ( $\chi^2(2) = 3.62, p = .16$ ) and no interaction ( $\chi^2(4) = 2.59, p = .63$ ). Post-hoc tests show that CM is rated significantly lower than AF ( $p = .0008$ ) and AM ( $p = .025$ ).

The direct trust ratings (Fig. 6), show the common pattern with a main effect for Identity ( $\chi^2(2) = 20.33, p < .0001$ ), but not a main effect for Style ( $\chi^2(2) = 4.11, p = .13$ ) and no significant interaction ( $\chi^2(4) = 2.78, p = .59$ ). Post-hoc results show that trust was significantly lower for CM than AF ( $p = .0001$ ) with a tendency to be lower than AM ( $p = .053$ ). The difference between AF and AM was not significant ( $p = .11$ ).

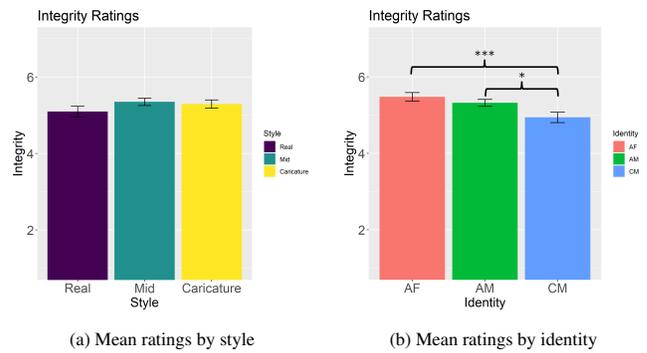


Figure 5: Integrity ratings, grouped by Style (a) and Identity (b) with lines pointing to significantly different pairs. \*\*\* indicates a  $p$ -value less than 0.001, and \* a  $p$ -value < 0.05.

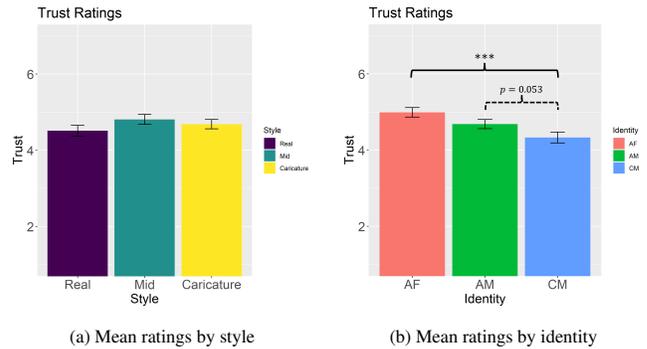


Figure 6: Trust ratings, grouped by Style (a) and Identity (b) with lines pointing to significantly different pairs. \*\*\* indicates a  $p$ -value less than 0.001.

## 6.2 Doctor Selection

At the end of the task, participants selected one doctor to get an opinion from (their most preferred doctor) and also selected their least preferred doctor. Figure 7 (a), (b), and (c) show all preference results. As with the survey results, there is a clear disinclination towards the CM identity, evidenced in Figure 7(b). Unlike with the survey results, however, there is also a style preference in the selections. The Exact Multinomial Test was used to test if variation in doctor selections based on Style was significantly different from chance, and binomial tests on each proportion with Bonferroni corrected  $p$ -values were used for post-hoc testing when necessary. We consider selection by chance to mean that each style was equally likely to be selected. We found that doctor selection by Style was unlikely to occur by chance ( $p = 0.044$ ), with the probability of Mid being the style of the most preferred doctor significantly higher than chance ( $p < 0.05$ ). In Figure 7(c), this can be seen for all three identities, with the preference for Mid most apparent for AF and CM.

Participants also selected their least preferred doctor. The Exact Multinomial test did not indicate that variation in the least preferred doctor by style was different from chance. Regarding identity, the responses largely mirror the results for the preferred doctor, with CM being selected as the least preferred most frequently and AF being selected as least preferred the least frequently (Figure 7(e)).

We do not see strong evidence that people were more likely to select doctors that matched their identity group. For example, the female doctor (AF) was selected by 44% of overall participants, 39% of those identifying as female and 50% of those identifying as male. The Caucasian male (CM) was selected by 15% of participants

overall, 18% of Asian participants (only 11 in this group) and 21% of Caucasian participants. CM was selected by only 4.5% of members of other racial groups (by 1 of 22).

## 7 STRUCTURED INTERVIEWS

After completing the doctor interviews and making their selection, all participants completed a semi-structured interview with the research assistant. These interviews consisted of five open ended questions, sometimes with additional questions to request clarifications. Answers were recorded, transcribed and then coded for analysis in order to look for common patterns in the responses.

The first question was “What made you select the doctor that you chose as the most preferred?”. In explaining their rationale, participants were very focused on the content of the interview as it related to qualifications and competency. This was mentioned by 44 out of 66 participants, with 29 mentioning experience or knowledge and 15 mentioning the quality of the answers in this regard. Twelve participants mentioned the doctors behavioral patterns, such as being focused on patient needs, dedicated to lifelong learning, or being willing to deliver negative news. In total, 50 participants mentioned these behaviors and/or content. Personality and affective traits were also important, but appeared secondary, being mentioned by 29 participants, with 24 referencing qualities like warmth and empathy. Identity features (e.g. age, gender, race) were mentioned by 15. Eight preferred a female doctor, five liked an older doctor. Nine participants mentioned principles or values. Only three participants mentioned avatar features, like realism or good facial motion.

The second question was “Why did you choose the doctor that you least preferred?”. Participant rationales fell in similar categories, but with different distributions. Content related to qualifications and competency was still important, but only mentioned by 28. Short or vague content was a rationale for 13 participants. Behavioral patterns, such as less commitment to thoroughness, were mentioned by an additional 3 participants. Personality and affective attributes were mentioned by 28, so on par with both content reasons and these attributes as rationale for selecting the most preferred doctor. Being less caring/warm (9), less trustworthy (7), too focused on him/herself (7), or less genuine (4) were all concerns. Identity factors were only mentioned by five, with three finding a doctor too young and two being less likely to trust a white male. Avatar and/or motion qualities were now mentioned by nine. Specific features tended to diverge, but included eye contact, poor appearance, a scary mouth, and not smiling. Values or principles were mentioned by seven. Interestingly, eight participants mentioned that they may have rated the first avatar that they saw lower because they were getting used to the experience at that point.

The next two questions were: “Did you notice any features about the avatars that stood out to you, either positive or negative?” and “What do you think of the appearance of each avatar and did that impact your selections at all?”. The content of the answers to these questions tended to overlap, so they were combined and coded as a single group.

The avatars were generally well received, with 21 participants categorizing them as fairly to very realistic and only four describing them as uncanny or creepy. Considering comments on individual avatars, an additional 4 thought CM looked creepy, but 2 people commented specifically on CM looking realistic.

Twenty two participants made positive comments about the quality of the motion compared to only four who thought it robotic or unnatural. Seven participants did mention noticing an issue with lip syncing on at least one avatar.

Eye contact and appearance generated the most comments, being mentioned by 34 participants. They were particularly aware what they perceived as good or bad eye contact, e.g. “[AF-Mid’s] eye movements were very realistic and very genuine. It didn’t feel robotic.” P54, as well as overly large eyes, e.g. “[AF-caricature

eyes]: Really large and just an unreal shape, really round. Rounder than any round I’ve ever seen, basically. That kind of round” P32 and “At first I was honestly a little bit off put by the first one [AF-caricature] just because the eyes were kind of weird, and also this is my first time doing anything like this so I didn’t really know what to expect.” P63.

The next most mentioned feature was the mouth, being mentioned by 14, this was dominated by 11 negative comments on CM’s lips, mouth shape, or teeth.

Twelve participants mentioned noticing avatar behavior, which came as a direct result of performance choices made by the actors.

Participants were aware of identity features, such as race, age and gender. These were mentioned by 18 participants. The age of AM was noted by seven. While AM is a South Asian male, four participants thought him African American and one Caucasian.

Participants thought were evenly split on whether the avatar impacted their decision, with 22 saying the avatar appearance or motion did not have an effect and 22 saying it did or might have.

While care was taken to recalibrate the floor height for each participant, seven indicated that the avatars appeared short or below eye level, indicating that there were some calibration issues.

The final question was “Did you prefer the overall style of one of the avatars over the others?”. Eighteen participants indicated no preference. An additional twelve did not provide a clear answer. Those stating a preference generally referred to a specific avatar. Fifteen suggested AM (seven of those had AM-real), 7 AF, and two CM. Grouped by style, 10 picked Real, 7 Mid, and 7 Caricature. Four people indicated that they preferred a more realistic avatar, but two people thought the mid avatar was the most realistic. Overall, there is little indication of a style preference, although some indication of AM being preferred overall. Interestingly, seven participants stated they thought the style of all the avatars was the same or similar. The research assistants also noted that some participants did not realize there were three intentionally different styles in the interaction.

## 8 FOLLOW UP COMPONENT TESTS

There is a clear preference against the CM identity in doctor selections. Given that *Identity* is a multifaceted factor in the study, consisting of the actor’s performance, the script, the appearance, and technical performance of the avatar, we wished to gain deeper insight into which aspects of identity led to the observed preference. We therefore ran follow-up studies online using Amazon Mechanical Turk to test components of the experience: audio-only, images of the avatars (image-only), and silent videos of the avatar responses.

### 8.1 Audio-only Experiment

The first study used only the actors’ audio recordings. This allowed us to isolate the actors’ vocal performances and the variations in the script as possible contributors to the observed difference. The scripts were pre-tested and found comparable (Sec. 3.2), so any variation here should result mostly from the vocal performance.

The study followed the same design, except that there was no *Style* factor. Everyone heard the same three doctor recordings used in the initial experiment, in randomized order, and completed the same surveys after each recording. After all three interviews, they selected their most and least preferred doctor. A total of 63 US workers on MTurk completed the study and were compensated at a prorated rate of \$15/hr. Given our own impression of the audio of CM as being less warm and genuine, we hypothesized that there would be a lower response to CM with comparable results to the full study.

The results did not support a significant preference or varied level of trust based on the voice performances alone. No statistically significant differences were found on any of the measures. Ratings were similar for the three trust components (Ability: AF M=5.96,

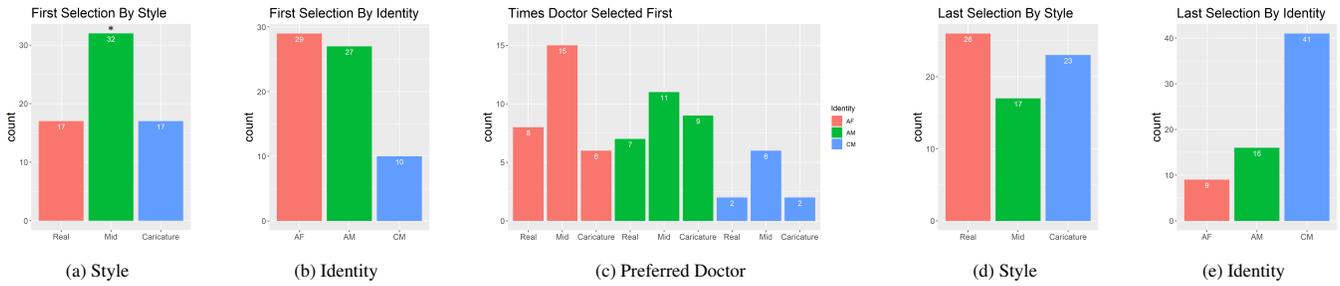


Figure 7: Preferred doctors, grouped by style (a) and identity (b). Graph (c) shows the frequency each doctor was selected as preferred, broken down by style. Graphs (d) and (e) show the least preferred doctors by style and identity. When grouped by style, Mid was the most frequently selected as preferred (a). \* signifies a  $p$ -value  $< 0.05$ .

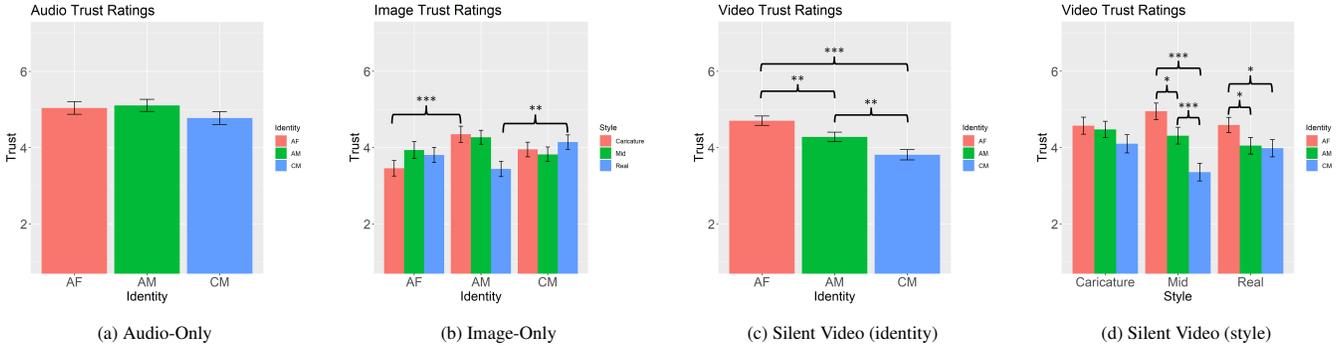


Figure 8: Trust ratings for the online studies using just the audio (a), just images of the avatars (b), and silent videos (c) & (d). These studies used a simplified study design and survey. Lines point to significantly different pairs. \*\*\* signifies a  $p$ -value  $< 0.001$  and \*\* a  $p$ -value  $< 0.01$ .

SD=0.92, AM M=6.15, SD=0.91, CM M=5.96, SD=0.89; Benevolence: AF M=5.77 SD=1.0; AM M=5.73 SD=1.0, CM M=5.66 SD=0.97; Integrity: AF M=5.66 SD=0.96, AM M=5.60 SD=1.1, CM M=5.48 SD=0.99). In terms of selection, CM was the least common first choice (AF 22, AM 23, CM 18), but was not the least preferred (AF 18, AM 24, CM 21), as in the original study. In both cases, these rather similar selection rates are significantly different.

The voice does not appear to account for the differences observed in the first experiment. These results also reinforce that script differences were not responsible for the avatar preferences.

## 8.2 Visual Comparisons

Two separate studies were run on Mechanical Turk to examine visual aspects of the experience. The first used only the images shown in Figure 2 and the second used silent videos of the response to the first question. These studies did not include dialogue or any additional information about the doctors, so the “interview” design of the main experience was unsuitable. Instead, we employed a simplified study design. Participants were told that they would see avatars of doctors from a new telehealth system. Each of the nine identity-style combinations were shown in random order and participants had to rate their agreement with the following statement: “I am likely to find this avatar trustworthy.” on a 7-point Likert scale. We collected responses from 63 MTurk workers for the image only study and 61 workers for the silent video study.

Results from the image only study showed no significant main effect for Style ( $\chi^2(2) = 2.47, p = .29$ ) nor Identity ( $\chi^2(2) = 5.14, p = .077$ ), but a significant interaction ( $\chi^2(4) = 22.2, p = .00018$ ). Post-hoc results show that AF Caricature was rated less trustworthy than AM Caricature ( $p = .0005$ ) and AM Real was rated less trustworthy than CM Real ( $p = .009$ ). These results do not show a similar pattern to the main study and are isolated to particular

avatars.

Results from the silent video study show no significant main effect for Style ( $\chi^2(2) = 2.5, p = .29$ ), but a significant main effect for Identity ( $\chi^2(2) = 48.9, p < .0001$ ) and a significant interaction ( $\chi^2(4) = 17.9, p = .0012$ ). Post-hoc analysis shows that there is a three level effect for Identity, with CM lower than AF ( $p < .0001$ ) and AM ( $p = .0010$ ) and AM less trustworthy than AF ( $p = .0028$ ). Figure 8(c) shows the post-hoc results for Identity. The interactions reflect the overall main effect for the Mid stylization and are shown in Figure 8(d). For Real, it is supported ( $AF > AM, p = .0408, AF > CM, p = .018$ ) except AM and CM Real ratings do not statistically differ ( $p = .95$ ). For Caricatures, there was a tendency for CM to be less trustworthy than AF Caricature ( $p = .084$ ), but other pairs were not significantly different. The overall identity trend is consistent with the main study, but this does not hold for particular Style-Identity pairs.

## 9 DISCUSSION

With regard to *Style*, we do not see evidence that people are more likely to trust more realistic avatars nor less likely to trust our most stylized caricatures. This runs counter to some previous work that suggests a preference for realistic avatars [18, 25], but is consistent with other work that found no such preference [5, 29]. Indeed, the strongest result related to style shows that people were more likely to select the doctor portrayed in the Mid-style, doing this 48% of the time, compared to 26% for the other two categories. It may be that this Mid-style avoided details that people found uncomfortable in the other two styles, perhaps uncanniness in the Real avatars or overly exaggerated features in the Caricature examples. Another explanation is that the match between avatar and motion may have been best at this style level. The motion from a different performer may not sufficiently match the avatar or have sufficient detail for

Real. There also might not be the level of exaggeration and motion stylization needed for Caricature to look correct. Other work found appeal rated highest for avatars between highly stylized and highly realistic [12]. The statistical power of the selection test is limited and it just passed our alpha cutoff ( $p = 0.044$ ), so while the study provides useful evidence, more work is likely required in order to draw strong conclusions for a Mid-style preference. It is also important to note that this style preference was not reflected in the trust survey ratings. Nonetheless, the study offers some evidence that semi-realistic avatars could be an effective choice even for sensitive applications where trust is important and no evidence that realistic avatars, at the level and with the scenario used here, improve trust.

The strongest and most consistent results relate to the *Identity* factor. There is a clear overall dispreference for the CM avatar. This is seen on the trust surveys, where CM is rated significantly lower than AF and AM on Ability, Benevolence, and Integrity, the constituents of trust in the model employed here. For the direct trust rating, he was seen as significantly less trustworthy than AF, and rated lower than AM, though not significantly ( $p = .053$ ). This dispreference is also seen in the doctor selections, where only 15% of participants selected CM (where 33% would be an even distribution) and 62% indicated that he was their last choice.

While Identity is a complex factor, most components of it can be ruled out as causing the observed preference. In the structured interview, people were most likely to indicate that they chose a doctor based on the content of their interview answers, with affective reasons coming second. It would thus be easy to interpret the identity preference to be based on the content of CM's answers being worse. However, this conclusion is not supported by either the pre-study of the scripts, which showed similar ratings on all selected dialogue, or the audio-only study, which also showed no significant differences across dialogue. The direct trust measure was slightly lower on the audio-only test, although below significance. Any contribution from audio or text appears small compared to the pronounced differences seen in the main study. Something other than content appears responsible. It is worth noting that although we had a diverse range of identities, we included only three in our study due to practical limitations. Perhaps including more options would help reveal what aspects of Identity influences preference.

The trust ratings on avatar images are similar across identities, so it seems like neither the overall look of the avatar nor group identity factors such as gender or ethnicity are driving the effect. The silent video ratings alone mirror the differences found in the main study. This suggests that the identity effect may be driven by either how well the avatar conveys the actors motions, on body and face, or perhaps by the quality of the actors body language itself. Both of these factors are present in both the main study and the silent videos. Each actor had unique non-verbal behavior, such as body language and gaze behavior, which may have influenced how the personality of the avatars was perceived [26]. Mapping between the actors and the avatars could also introduce error. There were also some issues with the avatars. For example, people commented negatively on the mouth of CM, yet were still likely to select this avatar for the Mid style, where this issue seems worse. There is not perfect alignment between the silent movies and full results as the significant differences for the interactions seen in Figure 8(d) were not seen in the main experiment.

It should be noted that the effects also cannot be reduced to visual appeal alone. A visual appeal survey on the images showed that CM-Real was rated highest and AF was generally lower, which is not consistent with the trust findings.

There are important considerations related to the stylization results. Even our most stylized caricatures are still reasonably realistic, so much so that seven participants commented on not noticing a style difference across avatars, something we thought was quite clear. These findings may not extend to lower fidelity, less realistic avatars

and the study does not test nonhuman or other forms of avatars. On the realistic end, while the motion and visual quality were quite high, it remains very difficult to create photorealistic avatars that move as naturally as real people. We no doubt remain short of that bar. The tendency for the Real avatars to be rated lower on Benevolence may suggest an impact of the uncanny. It is possible that results could change for avatars that are indistinguishable from actual people.

## 10 CONCLUSION

This study used a high stakes medical scenario in order to explore how people's tendency to trust was impacted by avatar stylization. Stylization had no significant impact on trust ratings. The Realistic avatars received lower ratings for benevolence on average; however, the differences were not significant. This suggests that realistic avatars are not required and it may be possible to use a range of avatar visual styles for applications where trust is critical. Such a conclusion needs to be moderated with the acknowledgment that the Real avatars were still short of perfectly realistic and results might change with that condition. People were more likely to select the Mid stylized doctor, despite this preference not being reflected in the subjective trust surveys. A possible explanation for this is that there were perhaps details in the more extreme avatars that made people uncomfortable, perhaps some uncanniness for the Real avatars and too much exaggeration for the Caricature. Another factor might be that the animation itself, which was motion captured, was not appropriate for the caricatured avatars. Though the animation was high fidelity while the doctors were speaking, their idle motion was limited, and was noted by one of the participants. These details would be softened in the Mid style. The overall motion quality was high throughout and the impact of stylization might increase if motion fidelity is lower. For instance, the nuanced, realistic movement might make even a caricatured model look human, but this might not be the case for lower quality motion.

Results suggest that avatar identity had a strong impact on trust, with one of the three identities being trusted less based on subjective surveys and also selected less for a second opinion. This is despite careful efforts to try to balance the content of all three avatars and a lack of evidence that content varied in meaningful ways.

There is still much future work to be done. It would be interesting to try to apply the motion from an actor to different avatars to see if the motion quality had the same impact as the avatar changed. This does introduce challenging retargeting issues. The scenario here supported limited interaction where the avatars could answer preset questions, but did not support full, free form conversation. It would be interesting to extend the work in that way, perhaps by using a live confederate to drive the avatar. It would also be useful to do a more fine tuned breakdown of different avatar features to try to understand which are important. Finally, it would be useful to test more extreme avatars. The caricatures appear to perform fine in this situation, but these were still relatively realistic humans. What if the avatar is a fish or a dinosaur? Does this break the connection? Establishing unacceptable conditions would allow for a more careful investigation of the representational requirements. Another worthwhile future direction would be to compare avatars directly to face-to-face human interaction.

This work suggests that a range of styles may be applicable in even high trust settings and offers some evidence that a semi-realistic style may be preferable. It also provides evidence that motion issues, either stemming from actor body language or avatar design, can impact trust.

## ACKNOWLEDGMENTS

We gratefully acknowledge the contributions of Alex Laurant, Aaron Ferguson, Vivian Lo, Callie Waldman and Zachary Duwenhoegger, as well as our avatar models.

## REFERENCES

- [1] H. M. Aljaroodi, M. T. Adam, T. Teubner, and R. Chiong. Understanding the importance of cultural appropriateness for user interface design: an avatar study. *ACM Transactions on Computer-Human Interaction*, 29(6):1–27, 2023. 2
- [2] D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*, 2014. 5
- [3] N. Bos, J. Olson, D. Gergle, G. Olson, and Z. Wright. Effects of four computer-mediated communications channels on trust development. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 135–140. ACM, 2002. 2
- [4] Z. Dai and K. F. MacDorman. The doctor’s digital double: how warmth, competence, and animation promote adherence intention. *PeerJ Computer Science*, 4:e168, 2018. 2
- [5] Z. Dai and K. F. MacDorman. Creepy, but persuasive: In a virtual consultation, physician bedside manner, rather than the uncanny valley, predicts adherence. *front. Virtual Real. 2: 739038. doi: 10.3389/frvir*, 2021. 2, 8
- [6] F. Danieau, T. Lopez, N. Mollet, B. Leroy, O. Dumas, and J.-F. Vial. Enabling embodiment and interaction in omnidirectional videos. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 697–702. IEEE, 2017. 2
- [7] P. Debevec, T. Hawkins, C. Tchou, H.-P. Duiker, W. Sarokin, and M. Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pp. 145–156, 2000. 4
- [8] Y. Ferstl, E. Kokkinara, and R. McDonnell. Do i trust you, abstract creature? a study on personality perception of abstract virtual faces. In *Proceedings of the ACM Symposium on Applied Perception*, pp. 39–43, 2016. 2
- [9] Y. Ferstl and R. McDonnell. A perceptual study on the manipulation of facial features for trait portrayal in virtual agents. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pp. 281–288, 2018. 2
- [10] Y. Ferstl, M. McKay, and R. McDonnell. Facial feature manipulation for trait portrayal in realistic and cartoon-rendered characters. *ACM Transactions on Applied Perception (TAP)*, 18(4):1–8, 2021. 2
- [11] Y. Ferstl, S. Thomas, C. Guiard, C. Ennis, and R. McDonnell. Human or robot? investigating voice, appearance and gesture motion realism of conversational social agents. In *Proceedings of the 21st ACM international conference on intelligent virtual agents*, pp. 76–83, 2021. 2
- [12] R. Fleming, B. J. Mohler, J. Romero, M. J. Black, and M. Breidt. Appealing female avatars from 3d body scans: Perceptual effects of stylization. In *VISIGRAPP (I: GRAPP)*, pp. 335–345, 2016. 2, 9
- [13] R. Fribourg, F. Argelaguet, A. Lécuyer, and L. Hoyet. Avatar and sense of embodiment: Studying the relative preference between appearance, control and point of view. *IEEE transactions on visualization and computer graphics*, 26(5):2062–2072, 2020. 2
- [14] G. Gorisse, O. Christmann, S. Houzangbe, and S. Richir. From robot to virtual doppelgänger: Impact of visual fidelity of avatars controlled in third-person perspective on embodiment and behavior in immersive virtual environments. *Frontiers in Robotics and AI*, p. 8, 2019. 2
- [15] F. Herrera and J. N. Bailenson. Virtual reality perspective-taking at scale: Effect of avatar representation, choice, and head movement on prosocial behaviors. *new media & society*, 23(8):2189–2209, 2021. 2
- [16] L. Hossain and R. T. Wigand. Ict enabled virtual collaboration through trust. *Journal of Computer-Mediated Communication*, 10(1):JCMC1014, 2004. 2
- [17] K. M. Inkpen and M. Sedlins. Me and my avatar: exploring users’ comfort with avatars for workplace communication. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pp. 383–386, 2011. 2
- [18] D. Jo, K.-H. Kim, and G. J. Kim. Effects of avatar and background types on users’ co-presence and trust for mixed reality-based teleconference systems. In *In Proceedings the 30th Conference on Computer Animation and Social Agents*, pp. 27–36, 2017. 2, 8
- [19] E. Kokkinara and R. McDonnell. Animation realism affects perceived character appeal of a self-virtual face. In *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games*, pp. 221–226, 2015. 2
- [20] M. E. Latoschik, D. Roth, D. Gall, J. Achenbach, T. Waltemate, and M. Botsch. The effect of avatar realism in immersive social virtual realities. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, pp. 1–10, 2017. 2
- [21] J.-L. Lugin, J. Latt, and M. E. Latoschik. Avatar anthropomorphism and illusion of body ownership in vr. In *2015 IEEE Virtual Reality (VR)*, pp. 229–230. IEEE, 2015. 2
- [22] F. Ma and X. Pan. Visual fidelity effects on expressive self-avatar in virtual reality: First impressions matter. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 57–65. IEEE, 2022. 2
- [23] R. C. Mayer and J. H. Davis. The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of applied psychology*, 84(1):123, 1999. 5, 12
- [24] R. C. Mayer, J. H. Davis, and F. D. Schoorman. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734, 1995. 1, 3
- [25] R. McDonnell, M. Breidt, and H. H. Bühlhoff. Render me real? investigating the effect of render style on the perception of animated virtual humans. *ACM Transactions on Graphics (TOG)*, 31(4):1–11, 2012. 2, 8
- [26] A. Normoyle, J. B. Badler, T. Fan, N. I. Badler, V. J. Cassol, and S. R. Musse. Evaluating perceived trust from procedurally animated gaze. In *Proceedings of Motion on Games*, MIG ’13, p. 141–148, 2013. 2, 9
- [27] S. Oberdörfer, D. Schraudt, and M. E. Latoschik. Embodied gambling—investigating the influence of level of embodiment, avatar appearance, and virtual environment design on an online vr slot machine. *Frontiers in Virtual Reality*, 3:828553, 2022. 2
- [28] N. Ogawa, T. Narumi, H. Kuzuoka, and M. Hirose. Do you feel like passing through walls?: Effect of self-avatar appearance on facilitating realistic behavior in virtual environments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2020. 2
- [29] H. Patel and K. F. MacDorman. Sending an avatar to do a human’s job: Compliance with authority persists despite the uncanny valley. *Presence*, 24(1):1–23, 2015. 2, 8
- [30] Y. Patotskaya, L. Hoyet, A.-H. Olivier, J. Pettré, and K. Zibrek. Avoiding virtual humans in a constrained environment: Exploration of novel behavioural measures. *Computers & Graphics*, 110:162–172, 2023. 2
- [31] R. Radiah, D. Roth, F. Alt, and Y. Abdelrahman. The influence of avatar personalization on emotions in vr. *Multimodal Technologies and Interaction*, 7(4):38, 2023. 2
- [32] L. Ring, D. Utami, and T. Bickmore. The right agent for the job? the effects of agent visual appearance on task domain. In *Intelligent Virtual Agents: 14th International Conference, IVA 2014, Boston, MA, USA, August 27-29, 2014. Proceedings 14*, pp. 374–384. Springer, 2014. 2
- [33] K. Ruhland, K. Zibrek, and R. McDonnell. Perception of personality through eye gaze of realistic and cartoon models. In *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception*, pp. 19–23, 2015. 2
- [34] M. Slater and M. Usoh. The influence of a virtual body on presence in immersive virtual environments. In *VR 93, Virtual Reality International, Proceedings of the Third Annual Conference on Virtual Reality*, pp. 34–42, 1993. 1
- [35] H. J. Smith and M. Neff. Communication behavior in embodied virtual reality. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2018. 1
- [36] S. Ter Stal, L. L. Kramer, M. Tabak, H. op den Akker, and H. Hermens. Design features of embodied conversational agents in ehealth: a literature review. *International Journal of Human-Computer Studies*, 138:102409, 2020. 2
- [37] I. Torre, E. Carrigan, K. McCabe, R. McDonnell, and N. Harte. Survival at the museum: A cooperation experiment with emotionally expressive virtual characters. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pp. 423–427, 2018. 2
- [38] O. Vereschak, G. Bailly, and B. Caramiaux. How to evaluate trust in ai-assisted decision making? a survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–39, 2021. 3

- [39] M. Volante, S. V. Babu, H. Chaturvedi, N. Newsome, E. Ebrahimi, T. Roy, S. B. Daily, and T. Fasolino. Effects of virtual human appearance fidelity on emotion contagion in affective inter-personal simulations. *IEEE transactions on visualization and computer graphics*, 22(4):1326–1335, 2016. 2
- [40] T. Waltemate, D. Gall, D. Roth, M. Botsch, and M. E. Latoschik. The impact of avatar personalization and immersion on virtual body ownership, presence, and emotional response. *IEEE transactions on visualization and computer graphics*, 24(4):1643–1652, 2018. 2
- [41] F. Weidner, G. Boettcher, S. A. Arboleda, C. Diao, L. Sinani, C. Kunert, C. Gerhardt, W. Broll, and A. Raake. A systematic review on the visualization of avatars and agents in ar & vr displayed using head-mounted displays. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 1, 2
- [42] P. Wisessing, K. Zibrek, D. W. Cunningham, J. Dingliana, and R. McDonnell. Enlighten me: Importance of brightness and shadow for character emotion and appeal. *ACM Transactions on Graphics (TOG)*, 39(3):1–12, 2020. 2
- [43] E. Zell, C. Aliaga, A. Jarabo, K. Zibrek, D. Gutierrez, R. McDonnell, and M. Botsch. To stylize or not to stylize? the effect of shape and material stylization on the perception of computer-generated faces. *ACM Transactions on Graphics (TOG)*, 34(6):1–12, 2015. 2
- [44] K. Zibrek, E. Kokkinara, and R. McDonnell. The effect of realistic appearance of virtual characters in immersive environments-does the character’s personality play a role? *IEEE transactions on visualization and computer graphics*, 24(4):1681–1690, 2018. 2
- [45] K. Zibrek, S. Martin, and R. McDonnell. Is photorealism important for perception of expressive virtual humans in virtual reality? *ACM Transactions on Applied Perception (TAP)*, 16(3):1–19, 2019. 2
- [46] K. Zibrek and R. McDonnell. Does render style affect perception of personality in virtual humans? In *Proceedings of the ACM symposium on applied perception*, pp. 111–115, 2014. 2
- [47] K. Zibrek and R. McDonnell. Social presence and place illusion are affected by photorealism in embodied vr. In *Motion, interaction and games*, pp. 1–7. 2019. 2

## SUPPLEMENTAL

### 11 SURVEYS

#### 11.1 Post-doctor Surveys

These surveys were completed after each doctor was seen. Participants were given the instruction “Think about the doctor you just interviewed. For each statement, select the number that best describes how much you agree or disagree with each statement.” All survey prompts were rated on 7-point Likert scales, with labels “Disagree Strongly, Disagree, Neither agree nor disagree, Agree, Agree Strongly.” The subsurvey titles (e.g. “Ability”) were not shown to participants, only the prompts. These surveys are based on those proposed by Mayer and Davis [23], with some customization to the presented medical scenario.

##### Ability

1. The doctor is very capable of performing his/her job.
2. The doctor has much knowledge about the work that needs done.
3. I feel very confident about the doctor’s skills.
4. The doctor has specialized capabilities that can increase performance.
5. The doctor is well qualified.

##### Benevolence

1. The doctor is very concerned about my welfare.
2. My needs and desires are very important to the doctor.
3. The doctor would not knowingly do anything to hurt me.
4. The doctor really looks out for what is important to me.
5. The doctor will go out of his/her way to help me.

##### Integrity

- 1.
2. The doctor has a strong sense of justice.
3. I never have to wonder whether the doctor will stick to his/her word.
4. The doctor tries hard to be fair in dealings with others.
5. I like the doctor’s values.
6. Sound principles seem to guide the doctor’s behavior.

##### Trust

1. If I had my way, I wouldn’t let the doctor have any influence over issues that are important to me.\*
2. I would be willing to let the doctor have complete control over my future healthcare.
3. I really wish I had a good way to keep an eye on the doctor.\*
4. I would be comfortable giving the doctor a task or problem which was critical to me, even if I could not monitor their actions.

\*-Reverse-scored item.

#### 11.2 Propensity to Trust Survey

This survey was conducted once, at the beginning of the experiment, and measures an individuals propensity to trust others.

##### Propensity

1. One should be very cautious with strangers.
2. Most experts tell the truth about the limits of their knowledge. Most people can be counted on to do what they say they will do.
3. These days, you must be alert or someone is likely to take advantage of you.
4. Most salespeople are honest in describing their products.
5. Most repair people will not overcharge people who are ignorant of their specialty.
6. Most people answer public opinion polls honestly.
7. Most adults are competent at their jobs.

### 12 TESTED SCENARIOS

In an online survey, participants were asked to “Please consider the following scenario:” and then respond to two rating prompts:

1. I can imagine what it would be like to be in this scenario.
2. Imagining myself in this scenario makes me feel vulnerable.

Each scenario is listed below.

1. Forensic Accountant / audit: You are a small business owner and have worked very hard for many years to build your business. You recently received a notice that you will be audited for not properly paying taxes. You suspect your former partner may have embezzled money that should have been used to pay taxes. You need to hire a forensic accountant to help you prepare for the audit. You will need to reveal a great deal of personal information to this person, so it is important that you find someone you can trust. If things go badly, you may lose your business.

2. Medical: You have been feeling off lately, and can’t figure out why. You decide to have short consultations with several doctors during which you describe your symptoms and ask for treatment advice. Each doctor offers differing advice, and it is up to you to decide which doctor you trust the most.

3. Medical kidney: Your doctor found a lump growing in your kidney. He says it could be benign, but it might be cancerous. You might even need to have the kidney removed. Overall, he seemed very uncertain and you want to get a second opinion before deciding on a course of action. Your life may depend on who you choose to get advice from, so you must choose a doctor you trust.

4. Lottery winnings: You have just won \$10 million dollars in the lottery and need to hire a financial manager to help you decide how to invest it. Your future wealth might depend heavily on the advice from the financial manager, so you must hire one that you can trust. If things go badly, you could lose all of your winnings.

5. Child Cancer: Please consider the following scenario: Your child has recently been diagnosed with cancer. Her current doctor does not have the resources to treat patients with cancer, so you must find a new doctor to provide her with medical care. Your child’s life may be at risk, so finding a doctor you can trust is of utmost importance. If the cancer is not treated properly, it could be fatal.

## 13 DIALOGUE

### 13.1 Ability

D12 (assigned to AF): I had the good fortune to attend the top school in the country for urologic oncology. I deal with patients with cancers anywhere in the urinary tract, but my main focus is on the kidneys. We've helped hundreds of patients. Now, I wish we could say we were 100% successful, but cancer is a nasty foe. Nonetheless, you'll be in good hands.

D9 (assigned to AM): I recently celebrated my 25th year of having admitting privileges at the top two hospitals in town. I've had the great privilege to serve an enormous number of patients with kidney issues during that time. There is no substitute for experience. I have seen so many different cases with unusual presentations of symptoms that go beyond the textbook. You'll be in good hands.

D7 (assigned to CM): That's interesting. Patients seldom ask that despite how important it is. First off, it is important that you always find someone that is trained in the correct specialty. I'm a nephrologist, so I specialize in treating kidney ailments. Second, and people rarely think of this, you need someone whose knowledge is current. I'm dedicated to a continuous education program where every year, I update my training.

### 13.2 Benevolence

D5 (assigned to AF): They're not really separable. I want to give my patients the best care that I can. That means that I must be at my best, so I need to get sleep and exercise and have time away to recharge. Then I'm fully able to give my all to my patients.

D3 (assigned to AM): It can be a challenge. I'm in medicine because I care about patients. I always want to do the best for them. I've learned, though, that I also need to take care of my own health or I'm of no use to anyone. It's a balance.

D14 (assigned to CM): It is important to remember that patients' are coming to me in a moment of crisis. Of course, I need to limit my working hours so that I don't burn out, but when I'm interacting with them, my patients' needs are always more pressing. It is easy and natural to prioritize them.

### 13.3 Integrity

D9 (assigned to AF): As a member of the American Medical Association, I'm committed to upholding a well formulated set of ethical standards. These include items like providing competent and compassionate care, following legal standards, and continuing to advance our own knowledge. When these become difficult in practice is when there appears to be a conflict. For example, I want to provide the best care I can to my patients, so I may want to talk to another doctor about the case, but I also need to protect the patient's confidentiality, which limits how I can share information. The answer is generally clear if you try to maximize protection for the patient, and there are ways to protect patient confidentiality and also maximize care.

D10 (assigned to AM): That is a thoughtful question. The most important principle that guides my work is that responsibility to my patient is paramount. I also always seek to operate in a professional manner, with compassion, and I try to stay current with the latest findings. A challenging situation arises when you have negative information that you know the patient will not want to hear. I consider it important to share these hard truths. The patient has the ultimate responsibility for making choices about their care, so they deserve to be fully informed.

D12 (assigned to CM): All the principles I follow are based on respect. Respect for the patient, respect for privacy, respect for the law, respect for the community. If I maintain this mindset, I find I am consistently acting in the best interests of my patients. The only situation I can think of where this can be a challenge is when I'm caught up in the rush of day to day tasks, I may get tired and look for something more expedient. In those moments, I just take a deep

breath and remind myself of the key principle of respect. That keeps me focused on taking the best care of my patients.

## 14 PARTICIPANT QUOTATIONS FROM STRUCTURED INTERVIEWS

Sample responses from Question 1, "What made you select the doctor that you chose as the most preferred?"

- "She just seemed really caring and the way she talked she seemed really well-qualified."P1
- "He focused on a more transparent relationship with the patient that doesn't place his power above that of the patient."P5
- "seemed really warm. He smiled a lot while he was talking to me. I think I have some bias because he appeared older. And I felt like values aligned with him."P3
- "Honestly, I think it's probably an inherent bias towards I like having female doctors. As a woman, I like being able to trust my healthcare to someone who I feel instinctively is going to understand more of what I'm physically going through"P7
- "Yeah, I would imagine, I would guess that I maybe was just paying closer attention and kind of respecting what he was saying more just because it was like oh, this looks more like a real person to me. That would be my best guess"P34
- "That was the one in the beginning [AF-low] I chose. I feel like she cared the most. The way she spoke was the most passionate and the most empathetic. You're treating me like a human I think most of all. She also seemed like she was not talking over me at all. The other 2 doctors, they both seemed like they were like "we're professionals. These question you're asking are dumb." She seemed much more honest and down to earth with me even though she was very honest about her own shortcomings, which I liked. I feel like that makes me trust her more. It just seemed like she gave a shit."P41
- "...She crafted a relationship with me based on her work and her respect of the medicine I think. Her language didn't feel self-aggrandizing..."P43
- "It was the first doctor [AM-med]. What made me choose him? [laughs] I got the instinctual feeling that the model of him, or the avatar or whatever, and the way he spoke, he was very Obama-ish. [laughs] He tends to give more eye contact in his responses. He portrayed himself to be very knowledgeable and had a good bedside manner."P50.

Sample responses to the question "Why did you choose the doctor that you least preferred?"

- "He [CM-med] didn't smile. If I'm really honest with myself, I think I had some attractiveness bias. His face looked kind of weird to me....He seemed very by the book, like I do the minimum and I do my job well ..."P3
- "He used a really big word at the very beginning- nephrology, I think. I was like, I don't know what nephrology is. So I wasn't sure whether or not he would be able to present things in a language that was accessible to me which would be important."P4
- "It's kind of Hannibal Lecter vibes off that one [CM.Real]. It's the thing where he was just kind of on the uncanny divide. There was a lot more detail in him that made him feel a little bit closer to being human visually a lot more in terms of the hair, the salt and pepper kind of thing. But then just thinking about

how his face moved. [[diff answer]] All of my answers for him are basically “he seems extremely capable and I wouldn’t trust him with a spoon.”P9

- “I could be wrong but I think this doctor focused a lot on having up to date academic knowledge. And I just think that that’s. . . It’s a very car salesman type of currency to brandish.”P5
- “I think he mentioned sometime if everything’s in a rush or something that he’d probably have a potential to lose control or something. And I just feel like that kind of attitude might make the patients doubt his experience or something. I personally did not expect that.”P30
- “I felt the doctor was very quick in answers which I felt dismissive. If I’m going through a hard time, I didn’t get a reciprocating feeling that they care. As interested as me as a person other than just the diagnosis.”P40
- “It would be the last one [CM-high], probably because I don’t really trust doctors and a lot of doctors that have done me dirty have been white males. His response about respect, I just didn’t buy it. Yeah it’s just an instinctual feeling. [laughs]”P50.

Sample responses for the questions: “Did you notice any features about the avatars that stood out to you, either positive or negative?” and “What do you think of the appearance of each avatar and did that impact your selections at all?”:

- “For instance, I think I remembered CM-med, you could see he was taking a second to think about what the question was. Yeah, just the fact that a lot of the body language of people when they’re explaining stuff was actually transferred which is interesting, but I guess that’s what makes people feel more comfortable is body language.”P2
- “There was a level of oh, this person is very human even though that’s not a human. But there’s a certain humanity embodied in that avatar. Like I said before, the eyes were really important. Their eyes and eye contact. I didn’t think of them like avatars, to be honest. I thought about them as people.”P4
- “I think the thing I noticed watching them was the slight lag in the movement of the faces and the talking. That was kind of the thing that popped up first in my brain of watching an avatar specifically”P7
- “I felt like I kind of towered over all of the avatars. And then also they look kind of little.”P25
- “Their expressions were pretty realistic. I can see how that would sway a person [inaudible], comforted by how they look or reactive by their facial expressions.”P27
- “Yeah, actually I was pretty impressed in many ways. I think the overall facial features and actions and also how they were moving their body was pretty accurate, and then there were parts that still felt very plastic just in terms of their smile or the way they were looking, like the direction they were looking. It kind of looked like they were...”P34
- “So let’s say I just finished clicking the first question and I’m about to click the second, the avatar just stands there, which I guess that’s what avatars do. But if it was a real doctor, the doctor would maybe go like this [demonstrates] or switch hand positions, something more natural. You never just stand and stare.”P36

- “His eyes were also too big. He looked like a weird demonic cartoon character. All of them were a little uncanny valley and I was creeped out by all 3 of them. I think it would be cool if they were just cartoons. If they were Miis from the Wii, I’d be more trustworthy of that than of these creepy faces plastered onto a PS2 avatar. So all 3 of them were scary in that way. The middle guy [AM-high], the older dude, he looked the most real. The hairline was very accurate looking. And so I think that might’ve been why I listened to his words I guess more. And the first one [AF-low] from the beginning, she wasn’t terrifying but all of them were unsettling.”P41
- “Her eye movements were very realistic and very genuine. It didn’t feel robotic. That being said, I also thought her mouth movements were not as good and the 2 male doctors, their mouth movements were much more realistic. However I kind of understood that they weren’t actually real so I don’t think the mouth movements impacted too much. But I think the eye movements definitely there was a sense of connection with the eyes being pretty realistic.”P54
- “I was more paying attention to what they were saying rather than what they were doing and physical appearances. I was just looking at their face when they were talking.”P1
- “They just seemed like people.”P2
- “The second guy [CM-low], when he first popped up I thought that maybe he looked a little bit creepy. But then all that he was saying and his demeanor and his voice was really reassuring, so that didn’t end up mattering much.”P8
- “In the beginning it was kind of uncomfortable. But then as it went on I felt like I could connect a little better with the doctors based on all the expressions on the face, those little things. That was pretty progressive towards the end. In the beginning it was a lot. [[additional answer]] I started to look into his eyes and then I could see the pupils kind of going back and forth like an actual person’s do. . . . But there were all kinds of little movements in his eyebrows and some of the skinfolds above his eyes, which kind of matched what he was saying. The script was trying to be sincere and then I think that came through with the facial expressions, too.”P10
- “Yeah, I tend to prefer female doctors just from my experiences with female organs and our experiences with being female. I personally prefer having female doctors because we know what’s going on in our bodies a little bit more. So yeah, the second female Asian lady [AF-low] was preferable too because of gender”P16
- “I think one of the reasons why I didn’t like the first [AM-low] and the third [CM-high], they seemed less real. They seemed more robotic. So I’m a salesperson also and I’m a people person, so for me it’s all about inflection and personality and the vibe I get from somebody. And so from both of them I got no vibe. I didn’t get a positive vibe from them. From her [AF-med] I got this nice, trusting, sweet but competent vibe. And I don’t know if it was because of their avatars, they didn’t really move very much. It was more in their tone than I think I picked that up. I think I was looking at trying to figure out what it was about. . . like why don’t I like this guy? You know what I mean? But I don’t know if it was because of the avatar or because of the tones of their voices.”P23
- “So I think in the way that the doctor I chose [CM-med], the way his face expressed empathy, it seemed to be the most trustworthy. . . . I think that the way his face expressed emotion worked really well”P29

- “The one thing that sticks out is the third doctor’s [AM-med] initial smile for some reason. It seemed like a good smile. It seemed like a professional smile. It was a practiced smile but it worked, if that makes sense.”P36
- “Yeah, I thought the African American guy [AM caricature] was great because that’s my demographic. I thought he was represented well. His features looked realistic. His hair looked almost digital kinky which was kind of cool.”P46
- “Well they looked pretty realistic.”P51
- “All 3 seemed very lean and very tailored. The hand coloring on the first man [one] who was a person of color, his hand coloring and texture did not match his face and his age. His hands actually almost looked like he was wearing gloves. So one would think that if he has a receding hairline and gray hair that they would’ve taken the time to match up the hands.”P56