

Effects of Nonsensical Responses in Virtual Human Simulations on Clinicians' Empathic Communication and Emotional Responses

Alexandre Gomes de Siqueira   [University of Florida | agomesdesiqueira@ufl.edu]

Heng Yao  [University of Florida | hengyao1993@ufl.edu]

Sarah Bloch-Elkouby  [Icahn School of Medicine | sarah.bloch-elkouby@mountsinai.org]

Megan L. Rogers  [Texas State University | megan.rogers@txstate.edu]


Olivia C. Lawrence  [Columbia University | ocl2108@tc.columbia.edu]

Devon Peterkin  [Columbia University | dpeterkin12@gmail.com]


Sifan Zheng  [University of Pittsburgh | sz2930@tc.columbia.edu]

Kathleen Feeney  [Florida International University | kfeeney@fiu.edu]

Erica D. Musser  [Barnard College | emusser@barnard.edu]

Igor Galynker  [Icahn School of Medicine | igalynke@gmail.com]

Benjamin Lok  [University of Florida | lok@cise.ufl.edu]

 Department of Computer and Information Science and Engineering, University of Florida, 432 Newell Dr, Gainesville, FL 32611, United States.

Received: 03 July 2024 • Accepted: 04 September 2024 • Published: 28 October 2024

Abstract In this manuscript, we report on research that explores the application of virtual human patients to train clinicians on empathic communication skills. During training, clinicians received empathy scores as they interacted with two virtual humans portraying suicidal ideation, who at times provided nonsensical responses. We video-recorded clinicians' interactions with virtual humans and analyzed their facial expressions, as well as their verbal responses. In phase I of our study, we analyzed clinicians' facial expressions during three key moments: after a sensical response from a virtual human (baseline), following the last nonsensical response of the interaction, and after a sensical response that followed the last nonsensical response. In phase I, facial expressions were grouped into Negative (anger, disgust, sadness, and fear) and Positive (happiness, neutral, and surprise) facial affective behaviors. We observed that nonsensical responses from virtual humans can negatively affect clinicians' positive and negative facial affective behaviors. We found a significant increase in the percentage of time clinicians express negative facial affective behaviors immediately following nonsensical responses. In phase II, we recruited additional clinician-participants and investigated how different proportions of nonsensical responses affect clinicians' facial expressions of individual basic emotions (instead of groups of positive and negative facial expressions), as well as whether nonsensical responses moderate the association between expressions of basic emotions and empathy scores obtained by clinicians during training. We observed a statistically significant positive interaction between proportions of nonsensical responses and angry facial expressions in predicting average empathy scores. That is, the relationship between anger and empathy scores was significant at low and mean levels of nonsensical responses, but not at high levels. These results suggest that at low and mean levels nonsensical responses negatively impact clinicians' performance, hindering their ability to acquire empathy skills. We discuss the impacts of technological limitations during virtual human interactions, particularly nonsensical responses, and the importance of controlling for such issues.

Keywords: virtual human; virtual patient; empathy; empathy skills; nonsensical responses; suicide crises syndrome;

1 Introduction

Virtual humans (VHs) have increasingly been employed to support humans in discussing uncomfortable and difficult topics. Researchers have sought to apply VH technology in communication skills training and health-related contexts [Rizzo *et al.*, 2011; Yao *et al.*, 2020]. For example, VHs have been used in interpersonal communication training between physicians and high-risk suicidal patients [Yao *et al.*, 2020], promoting colorectal cancer prevention with the public [Krieger *et al.*, 2021], supporting service members' Post Traumatic Stress Disorder treatment [Rizzo *et al.*, 2011]. Of-

ten, in training simulations, clinicians interact with standardized VH patients. Standardized VH patients can provide controllable, secure, and safe learning environments for repetitive practice [Stevens *et al.*, 2006]. However, while research investigating training contexts mostly consider *errorless* VHs, the current technology typically applied to build VH systems often make them prone to errors. For example, when natural language processing systems fail to interpret questions, or question-answer pairs are mismatched in the dialog model, VHs may give *nonsensical* responses [Skarbez *et al.*, 2011]. This may have profound implications for VH-based simulations given that people tend to treat com-

puters as social actors [Nass *et al.*, 1994] and expect VHs to respond as real humans would. Thus, it is important to investigate how humans respond verbally and non-verbally to VHs' mistakes, as well as how such mistakes impact the outcomes of VH-based training.

In this manuscript, we extend our research investigating communication errors during VH interactions. We report on our efforts investigating the effects of VHs' sensical and nonsensical responses on clinician-participants' facial expressions and overall performance during a simulation that seeks to train clinicians on empathic communication skills. Clinicians interact with standardized virtual patients represented by two VHs that demonstrate symptoms of the Suicide Crisis Syndrome (SCS) [Bloch-Elkouby *et al.*, 2021; Schuck *et al.*, 2019; Galynker, 2017]. SCS is associated with imminent suicidal behavior and characterized by pervasive feelings of entrapment in which escaping from an unbearable life situation is perceived as both urgent and impossible. The condition is also characterized by feelings of affective disturbance, loss of cognitive control, hyperarousal, and social withdrawal [Bloch-Elkouby *et al.*, 2020; Galynker, 2017]. During training, clinicians' verbal responses to empathic opportunities¹ were classified and scored using the Empathic Communication and Coding System (ECCS), a system designed to evaluate clinicians' empathy levels during medical interviews [Bylund and Makoul, 2002]. Nonsensical responses emerged during interactions, for example, due to issues in the conversational model or the speech-to-text conversion.

In phase I of our study [Gomes de Siqueira *et al.*, 2021], we analyzed the effects of VHs' sensical and nonsensical responses on clinicians' facial expressions during training. Due to the preliminary nature of this phase, the basic emotions elicited by the facial expressions were grouped into Positive and Negative. The Positive Facial Affective Behaviors (PFAB) group is composed of happiness, neutral, and surprise basic emotions, while the Negative Facial Affective Behaviors (NFAB) group includes anger, disgust, sadness, and fear emotions. The results suggest that nonsensical responses have a negative effect on clinicians' facial expressions during training and provide a broad understanding of the impact of nonsensical responses in clinician-VH interactions. However, these results do little toward eliciting how nonsensical responses affect individual facial expressions and clinicians' performance (ECCS scores) during the empathy training.

In phase II, with the addition of several participants, we analyzed how different levels of nonsensical responses affect clinicians' facial expressions, individual emotions, and performance during training. To this end, we examine whether different proportions of nonsensical responses during the clinician-VH interactions moderate the relationship between individual facial expressions (neutral, surprise, anger, happiness, sadness, disgust, and fear) and the clinicians' average ECCS scores. We hypothesize that exposure to different proportions of nonsensical vs sensical responses and the basic emotions clinicians demonstrate will negatively affect and differently predict ECCS scores obtained by clinicians during training. Taken together, our efforts aim to elicit

how a VH system that may provide nonsensical responses impacts clinicians' facial expressions (and underlying basic emotions) and how nonsensical responses impact training effectiveness. Broadly speaking, this work adds to the body of knowledge investigating how technologically motivated communication errors that occur while interacting with VH systems can impact participants' overall emotions and outcomes during training simulations.

2 Related Work

2.1 Technology-motivated Errors and Virtual Humans

Others have explored virtual agents in the context of technological errors [Lucas *et al.*, 2018; Wang *et al.*, 2013; Desai *et al.*, 2013; Skarbez *et al.*, 2011]. Lucas *et al.* [2018] explored whether culture impacts how social dialogue can mitigate the loss of trust caused when agents make conversational errors. They found that culture matters, suggesting a three-way interaction between culture, the presence of social dialogue, and the presence of errors [Lucas *et al.*, 2018]. Wang *et al.*'s work focused on the impact of errors in the ability of virtual agents to persuade users [Wang *et al.*, 2013]. Their system, similar to our own, made occasional mistakes such as giving irrelevant responses. Their results suggest that VHs who make mistakes during dialog are still capable of social influence, however, some users were more distracted by errors than others and overlooked the underlying persuasive information, rendering the simulation less effective. Desai *et al.* [2013] showed that users demonstrate trust loss if exposed to technological errors early in the interaction. They suggest that trust demonstrates inertia but may change over the duration of interactions. Skarbez *et al.* [2011] investigated conversational errors between participants and VHs. They analyzed the number of statements made by the participants and VHs, the number of errors, and the duration of the interaction. They propose a new method to evaluate VH interaction errors that were found to be significantly correlated with the Maastricht Assessment of Simulated Patients (MaSP) [Wind *et al.*, 2004]. Our work complements these efforts by exploring the relationship between conversational errors, participants' facial expressions (and underlying emotions), and performance during training.

2.2 Discussing Mental Health with Virtual Humans

Several research efforts employed VHs in mental health scenarios. VHs have been found to be invaluable tools to support humans acquire skills to have difficult conversations, including those related to mental health. Like our approach, SimSensei [DeVault *et al.*, 2014] examines users' verbal and nonverbal communication during their interaction with VHs. During the VH interaction, participants practice talking about mental health with a VH before talking to a real clinician. In another effort, eSMART-MH [Pinto *et al.*, 2015] focuses on helping users with depression communicate with clinicians by having them practice with VHs. The system

¹Empathic opportunities are critical moments when patients directly and clearly express their challenges, emotions, or progress to a clinician during an interview [Bylund and Makoul, 2005].

Table 1. Examples of nonsensical responses. Nonsensical responses were responses by VHS that had no meaning or made no sense in the context of the clinician-virtual patient dialog.

Participant's Question	Nonsensical Response
<i>What's your medical history?</i>	I am Cynthia.
<i>Have you taken any risks lately?</i>	I take something for my thyroid, and I'm on birth control.

has been found to be an effective tool to prepare users to talk to real clinicians about their health and to reduce users' anxiety toward sharing information about their issues. Several systems, such as ours, focus on supporting clinicians to acquire skills by interacting with VHS. Stuart *et al.* [2020] employed augmented reality to induce stress in nursing students to teach them stress management techniques using VHS. In another study, Stuart *et al.* [2022] explored learners' ability to identify verbal communication mistakes when interacting with VHS and how communication mistakes can affect their perceptions of credibility, reliability, and trustworthiness in the VH. Their work shows that learners overlook infrequent verbal communication mistakes and that mistakes may temporarily lower learners' credibility, reliability, and trust.

2.3 Virtual Humans and Empathy Training

Clinicians acknowledge that empathy skills are critical in forming therapeutic relationships based on trust and open communication with patients [Kim *et al.*, 2004]. Researchers have found that VH-based simulations can provide low-pressure environments to effectively train medical students' empathic communication skills [Kleinsmith *et al.*, 2015]. Others have observed that empathy training with VHS can increase clinicians' awareness of their verbal and nonverbal responses in situations in which patients disclose suicidal tendencies or intentions [Foster *et al.*, 2016]. It has also been demonstrated that VH simulations can be used to train clinicians' verbal and nonverbal skills to express empathy toward real patients [Foster *et al.*, 2016]. These works underscore the importance of VH training related to empathy skills acquisition. In this context, Empathic Opportunities have been specifically developed to help elicit empathic responses from users (typically clinicians) interacting with standardized VHS patients. [Borish *et al.*, 2014; Foster *et al.*, 2016]. The level of empathic responses elicited by empathic opportunities is typically evaluated based on the ECCS proposed by Bylund *et al.* [Bylund and Makoul, 2002; Foster *et al.*, 2016].

3 Study Context

This study is part of a broader, multi-year project aimed at training clinicians to effectively respond with empathy to virtual human patients who exhibit SCS [Schuck *et al.*, 2019; Galynker, 2017]. The overarching goal of the project is to enhance clinicians' ability to recognize and respond to empathic opportunities—critical moments during interactions when patients clearly express their emotions, progress, or challenges [Bylund and Makoul, 2005].

Clinicians participating in the study engaged with two VH patients demonstrating symptoms of SCS and a high risk for

near-term suicidal behavior. During these interactions, clinicians were trained and evaluated on their ability to provide high-empathy responses. Clinicians could ask questions verbally or type them out, and the VHS responded verbally, with lip-synced animations enhancing the realism of interactions.

However, the VH system, having undergone years of maintenance and adaptation by multiple research teams, sometimes produced nonsensical responses depending on the clinicians' questions (Table 1 shows examples of nonsensical responses). While these nonsensical responses could be seen as a limitation, they also contribute to the study's external validity. Similar flawed behaviors are likely to emerge in natural language processing systems over time, making this aspect of the study reflective of real-world scenarios.

Clinicians' empathy skills were assessed using ECCS, which categorizes and scores the level of empathy displayed during these key empathic opportunities. Empathy scores were computed based on clinicians' responses, with higher scores indicating a greater ability to engage empathetically with the VH patients. Previous research has shown that therapists' empathy skills are strong predictors of positive therapy outcomes [Ross and Watling, 2017], further underscoring the importance of this training. In our study, the ECCS was used to manually code and evaluate the clinicians' empathy levels during their interactions with the VHS, providing a critical measure of their ability to connect emotionally and supportively with patients.

4 System Description

Participants accessed the Virtual People Factory (VPF2) platform and interviewed a couple of standardized virtual patients. VPF2 is a VH authoring system [Rossen and Lok, 2012]. With VPF2, researchers can develop web-based, conversation-specific VH training simulations. VPF2 virtual humans can communicate verbally and nonverbally. Verbal communication is based on *conversational models* that allow VHS to hold conversations around predefined topics.

Conversational models consist of several tuples of multiple questions associated with a single response:

$$(n\text{-questions}, 1\text{-response})$$

This means that multiple questions from clinician-participants are associated with a single response from VHS. For example, if a clinician says, "Good morning," "Hi," or "Good evening," they may receive the same response from the VH: "Hi. I am glad to see you."

Conversational models may grow over time and become large (over 300 question-response tuples) as topic-specific conversational models are refined and extended, possibly across multiple studies and research efforts led by distinct

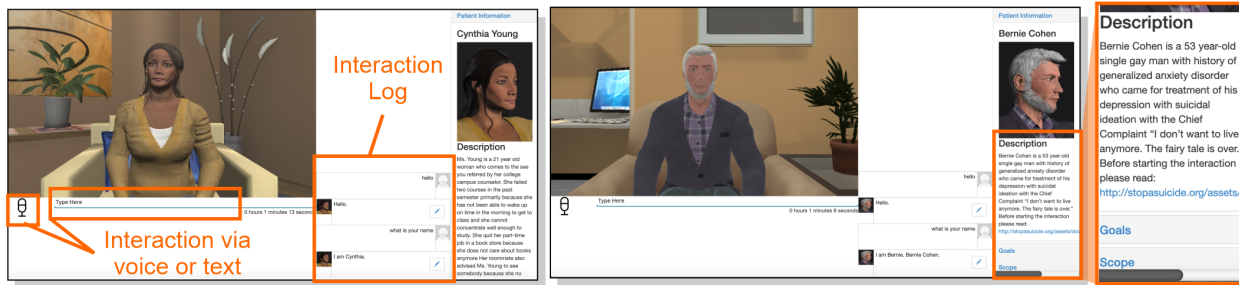


Figure 1. Virtual patients utilized in the study. (Left) Cynthia Young and (Right) Bernie Cohen. Both patients display pre-suicidal behavior and can discuss their symptoms and background stories with clinicians. Participants could interact via voice or text. Interactions were logged by the system. Participants could also access a description of the virtual patient, goals, and scope of the interaction using the right-side panel.

teams of researchers and practitioners. Maintenance and technological issues of such large conversational models may impact communication. Failure to customize conversational models adequately represents a source of nonsensical responses during VH interactions. Additionally, failure to adequately convert human speech to text, for example, due to limitations in the sound capture apparatus is another potential issue that may lead to nonsensical responses. In this study, nonsensical responses originated both from issues in the conversational model and speech-to-text technology.

The data collected in this study was obtained from the interactions of clinicians that accessed the VPF2 system with individual interaction web links sent to them by the research staff. Clinicians interacted with VPF2 VHS remotely, using their own equipment (laptop or desktop machines) or using equipment (Dell latitude 7290 laptops) provided to them in a lab setting. Figure 1 illustrates the VPF2 interface.

Clinicians were instructed to interact with the VHS preferably by voice. However, they could also use a keyboard to interact with the VHS. Preliminary data analysis suggested that clinicians that used the keyboard looked down (toward the keyboard) while typing, hindering our ability to properly capture their facial expressions. For this reason, for clinicians that typed their responses, a disproportionate percentage of their facial expressions were classified as “Unknown” by the facial expression recognition system. Additionally, it is possible that the mode of interaction (spoken vs. typed) could represent a confounding variable. Therefore, we only considered clinicians who consistently looked at the webcams and communicated verbally with VHS.

4.1 Virtual Patients Developed

The VHS Bernie Cohen and Cynthia Young were used in this study. Both VHS displayed pre-suicidal SCS symptoms. The 3D models of Cynthia and Bernie were developed using Adobe Fuse. Cynthia is capable of understanding 4310 questions and can provide 310 distinct answers, whereas Bernie’s conversational model allows him to respond to 2186 questions with 286 distinct answers. We employed graduate healthcare trainees to validate the questions-answer tuples of the conversational models. The trainees were supervised by our mental health expert collaborators.

Both VHS had specific backstories associated with them. Cynthia Young was a 21-year-old female college student who had failed two courses in the previous semester and was consistently arriving late to classes. She was referred to the clin-

ician by a college counselor. Additionally, she had trouble concentrating, which affected her ability to study. Bernie Cohen had a generalized anxiety disorder. As a 53-year-old gay man, Bernie was seeking treatment for depression and presented several SCS symptoms [Galynker, 2017].

VHS were displayed at the top-left corner of the user interface and occupied approximately two-thirds of the screen. The rest of the screen’s real state was dedicated to a chat box, information about the VH’s backstory, and buttons for navigation (See Figure 1).

4.2 Facial Expression Analyses

Clinician-participants video-recorded their facial expressions while interacting with the VHS. Later, the recordings were analyzed using the Noldus Facereader Automated Facial Coding (AFC) software [Lewinski et al., 2014]. AFC can recognize individual basic human emotions by analyzing facial expressions. First AFC creates a 3D Active Appearance Model (AAM) based on a clinician’s face [Cootes and Taylor, 2004]. Then, AAM obtains facial expressions’ intensity and probability on a continuous scale from 0 to 1.

AFC identifies seven basic emotions: happiness, sadness, anger, surprise, fear, disgust, and neutral [Ekman et al., 1969; Ekman and Cordaro, 2011]. AFC may also classify facial expressions as unknown. AFC has been validated and has shown high internal reliability [Cohen et al., 2013]. Based on the Noldus Facereader data, we calculate the percentage of time clinicians demonstrate each emotion in the video segments analyzed. In phase I of our study, facial expressions were grouped as Negative and Positive emotions. Sadness, anger, fear, and disgust were grouped as Negative Facial Affective Behaviors (NFABs), following the grouping adopted by [Cohen et al., 2013]. Toward accounting for all the emotions, the PFABs grouped the remaining facial expressions (Happiness, Surprise, and Neutral). We recognize that others have pointed out some ambiguity with respect to the valence of the Surprise emotion [Cohen et al., 2013]. This issue was not present in phase II, since each facial expression was analyzed individually.

4.3 Study Design and Procedure

This study has been approved by the Internal Review Board (IRB) and has followed the procedures in accordance with the ethical standards of the institutional committee on human experimentation. In this study, clinicians interacted with the

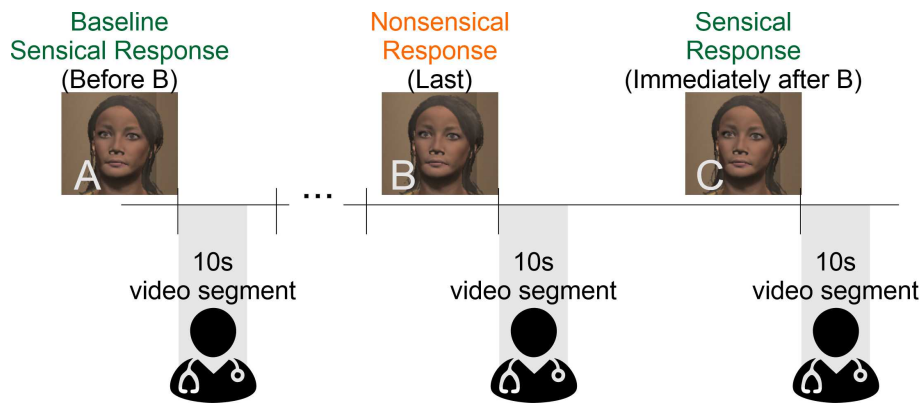


Figure 2. Phase I: Video Segmentation Strategy. Three ten-second segments were generated from the interactions of each participant that met the inclusion criteria (see Section 5.1.2). In total, thirty segments were analyzed.

system twice, with one week between each interaction. Once clinicians accessed the VH system, they were first introduced to the system's interface and how to interact with VHS. This was followed by information regarding empathic opportunities. Next, clinicians began interviewing the VH. First, clinicians interacted with Cynthia. During the interaction, they were free to interview Cynthia for as long as they wanted until they felt that they had enough information about the VH's condition. After interacting with Cynthia, clinicians accessed a feedback page with information regarding their responses during the empathic opportunities that emerged in the interview portion of the training simulation. After one week, clinicians followed a similar procedure, this time interacting with Bernie.

After interacting with the virtual patient in the second week, clinicians completed the "Virtual Human Interaction Satisfaction Survey" [Foster *et al.*, 2016], a self-report questionnaire designed to assess the feasibility, acceptability, and perceived utility of the virtual human interaction. The survey included 10 items, addressing various aspects of the VH interaction, such as whether the interaction helped clinicians learn how to formulate questions, whether the VH's answers were appropriate, and if the interaction simulated real life.

The responses and facial expressions analyzed were those recorded during the clinicians' interactions with Cynthia and Bernie in weeks one and two. We followed the strategy adopted by Yao *et al.* [2020] and combined the data regarding clinicians' interactions with Cynthia and Bernie since they were both developed to demonstrate symptoms of SCS, only differing in their appearance and background story. We compared participants' interactions with Bernie and Cynthia in relation to the duration and number of empathic opportunities elicited. It was observed that participants interacted with Bernie (mean=20.75 min, st. dev.=8.72 min) and Cynthia (mean=21.05 min, st. dev.=8.7 min) for about 20 minutes, triggering on average 18 empathic opportunities, Bernie (mean=18.1, st. dev.=8.77) and Cynthia (mean=18.65, st. dev.=9.12). This shows that clinicians' interactions with each of the VHS were similar regarding the duration and number of empathic opportunities triggered, supporting our approach to consider their combined interactions.

5 Phase I: The Effects of Nonsensical Responses on Participants' Grouped Facial Expressions

In the next sections, we briefly describe the methodology and results of phase I of this study [Gomes de Siqueira *et al.*, 2021], and how it motivated phase II.

Phase I examined the facial expressions of clinicians as they interacted with VHS who, at times, provided nonsensical responses. Clinicians' facial expressions were analyzed at three 10-second key moments during the interaction: moment 1) the baseline sensical response, moment 2) the last nonsensical response, and moment 3) the sensical response that occurred immediately following the last nonsensical response (See Figure 2). Thirty segments were analyzed.

We analyzed the percentage of time clinicians demonstrated NFABs and PFABs. We predicted that VHS' nonsensical responses would significantly affect clinicians' facial expressions. Our hypothesis was that we would observe an increased percentage of time clinicians demonstrated NFABs after nonsensical responses compared to after sensical responses. Conversely, we hypothesized that a sensical response that immediately followed a nonsensical response would counteract the effects of the nonsensical responses and reduce the percentage of time clinicians showed NFABs.

5.1 Methodology

5.1.1 Participants

When phase I was conducted, twenty clinician-participants (13 males and 7 females) had been recruited by the larger empathy research project this study is part of. Participants were psychiatry residents, clinical social workers, and psychologists. Participants were recruited from the Mount Sinai Health System (MSHS) and its three integrated hospitals, and the Florida International University (FIU) – affiliated with the Citrus Health Network. We recruited clinicians that had active caseloads, without experience with VH training systems. We did not use crowdsourcing to collect responses. Phase I was conducted in early 2020. Three of the participants interacted with the VHS remotely due to the COVID-19 pandemic using their own equipment. Ten participants met the inclusion criteria of the study (See Section 5.1.2). Re-

garding gender, two participants were females, and the rest were males. The age of participants ranged between 27 and 46 (mean= 31.8, st. dev. = 10.85).

5.1.2 Inclusion Criteria

These are the inclusion criteria of phase I:

- **Verbal responses:** VH's questions must have been answered verbally. The participant was not considered if responses were typed using a keyboard.
- **Number of empathic opportunities:** During the interaction, at least one empathic opportunity must have been triggered by the participant.
- **One sensical response immediately after a nonsensical one:** After an empathic opportunity, the participant must have received at least one sensical response that immediately followed a nonsensical one.

5.1.3 Segmentation of videos

The video recordings of clinicians interactions with VHs that met the inclusion criteria were segmented for analysis. For each recorded clinician-VH interview session, we generated three 10-second segments at specific moments during the interaction (See Figure 2).

- **Segment 1 – reaction to an empathic opportunity:** 10 seconds after the VH provided a sensical response to an empathic opportunity.
- **Segment 2 – reaction to the last nonsensical response from the VH:** 10 seconds immediately after the last nonsensical response of the interaction.
- **Segment 3 – reaction to a VH's sensical response that followed the last nonsensical one:** 10 seconds immediately after a sensical response from the VH that followed the last nonsensical response.

For the data analysis, we used statistical methods to analyze the percentage of time participants demonstrated the seven basic facial expressions identified in each of the video segments generated.

5.2 Results

In this section, a summary of the results containing only significant findings is presented. Our goal is to describe these findings from the perspective of motivating phase II of our study. For the complete analysis, please refer to Gomes de Siqueira et al. [2021].

We applied non-parametric statistical sign tests during data analysis. The sign test statistical method tests for differences between pairs of observations [Dixon and Mood, 1946]. We applied a non-parametric test because our data did not meet the equality of variances and normality of distributions assumptions. We used the sign tests to compare the percentage of time participants demonstrated PFABs and NFABs (See Figure 3). We conducted six tests, which we describe below (See Table 2).

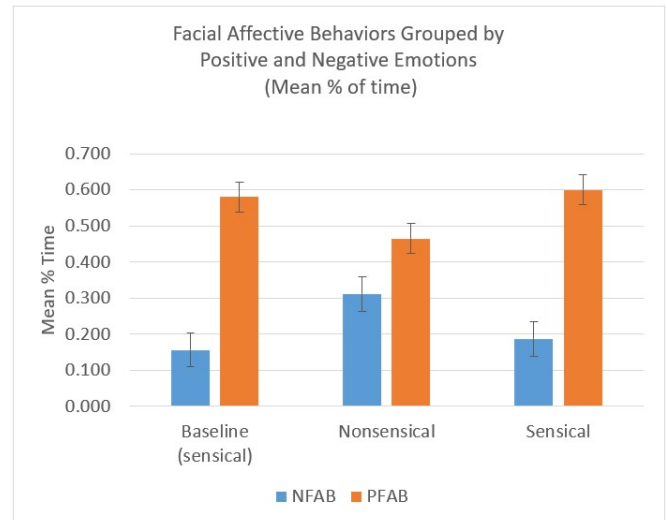


Figure 3. Facial Affective Behaviors' Mean Percentage of Time per Segment. Emotions are grouped by NFABs (negative) and PFABs (positive). Bars represent standard error.

1. **NFAB – nonsensical vs. baseline.** Comparing the percentage of time participants demonstrated NFABs after the last nonsensical response versus after the baseline sensical response.
2. **NFAB – nonsensical vs. sensical.** Comparing the percentage of time participants demonstrated NFABs after the nonsensical response versus after the subsequent sensical response.
3. **NFAB – baseline vs. sensical.** Comparing the percentage of time participants demonstrated NFABs after the baseline response versus after the subsequent sensical response.
4. **PFAB – nonsensical vs. baseline.** Comparing the percentage of time participants demonstrated PFABs after the nonsensical response versus after the baseline sensical response.
5. **PFAB – nonsensical vs. sensical.** Comparing the percentage of time participants demonstrated PFABs after the nonsensical response versus after the subsequent sensical response.
6. **PFAB – baseline vs. sensical.** Comparing the percentage of time participants demonstrated PFABs after the baseline response versus after the subsequent sensical response.

Next, we detail the significant results that were observed. Data are medians unless otherwise stated.

(1) **NFAB – nonsensical vs. baseline.** We found a statistically significant difference in the percentage of time clinicians demonstrated NFABs after nonsensical responses (mean= 0.311, st. dev= 0.317) compared to after baseline responses (mean= 0.156, st. dev= 0.255), $p = .031$.

(5) **PFAB – nonsensical vs. sensical.** We observed a statistically significant difference in the percentage of time clinicians demonstrated PFABs after nonsensical responses (mean= 0.465, st. dev= 0.353) compared to after sensical responses (mean= 0.6, st. dev= 0.36), $p = 0.039$. The results of the other tests were not significant.

Table 2. Facial Affective Behaviors Grouped by Negative and Positive Emotions (mean percentage of time.) Negative Facial Affective Behavior groups negative emotions (anger, disgust, sadness, and Fear), while Positive Facial Affective Behavior groups positive emotions (happiness, neutral, and surprise). Emotions not classified were labeled as “Unknown” and were not considered.

<i>Grouped Emotions</i>	Baseline Sensical		Nonsensical		Sensical	
	Mean (% time)	Std. Dev.	Mean (% time)	Std. Dev.	Mean (% time)	Std. Dev.
<i>NFAB</i>	0.156	0.255	0.311	0.317	0.187	0.304
<i>PFAB</i>	0.58	0.37	0.465	0.353	0.6	0.36

5.3 Discussion

Analyses 1 and 5 presented statistically significant results. Analysis 1 supports our hypothesis. We hypothesized that participants would demonstrate NFABs for an increased percentage of time after nonsensical responses compared to after sensical responses. In analysis 1, we observed that participants demonstrated NFABs for longer after nonsensical responses (mean= 0.311, st. dev= 0.317) when compared to after baseline (mean= 0.156, st. dev= 0.255) responses. Conversely, in analysis 5 we found an increase in the percentage of time participants demonstrate PFABs after a sensical response that immediately follows a nonsensical one (mean= 0.6, st. dev= 0.36) compared to after a nonsensical response (mean= 0.465, st. dev= 0.353). Following the terminology set by Desai *et al.* [2013], these results suggest that the effects of nonsensical responses on participants' facial expressions can be changed if participants are exposed to a sensical response that follows a nonsensical one.

Taken together, these findings suggest that VHs' nonsensical responses can have significant effects on participants' reactions and facial expressions during training. These results are relevant since facial behaviors play a critical role in clinician-patient interactions, and motivated us to further investigate the effects nonsensical responses during clinician-VH interactions. Next, we describe phase II of our study, which investigates the effects of nonsensical responses on each facial affective behavior identified independently, as well as clinicians' performance as measured by the empathy scores clinicians obtained during training.

6 Phase II: The Effects of Nonsensical Responses on Clinicians' Individual Facial Expressions and Empathy Scores

The results of Phase I provide a broad understanding of the impact of nonsensical responses on clinician-VH interactions since it analyzes facial expressions grouped by Negative and Positive emotions. However, it does little toward eliciting how the nonsensical responses and individual basic emotions affect overall clinicians' empathy skills acquisition during the training simulation and overall performance. Phase II of this study aims to address these limitations. To this end, phase II focuses on examining whether different proportions of nonsensical responses during the VH interactions moderate the relationship between individual facial expressions (neutral, surprise, anger, happiness, sadness, disgust, and

fear) and clinicians' average empathy scores obtained during training. We hypothesized that exposure to different proportions of nonsensical vs sensical responses during the training sessions and the different basic emotions clinicians demonstrate would differently predict empathy scores obtained by the clinician participants.

6.1 Methodology

The study design and procedure of phase II followed the steps described in Section 4.3. The data analysis of phase II was conducted in the early months of 2022 and considered all data available by the time the data analysis was performed. Compared to phase I, several additional participants were considered which allowed the analysis of individual emotions (instead of groups) and their impact on ECCS scores.

6.1.1 Participants

The additional clinician-participants considered in phase II were recruited from the same pool used in phase I, the Mount Sinai Health System and the Florida International University. Participants had to be psychiatry residents, social workers, or psychologists, have active caseloads, and have no prior experience with VH systems. By the time the data analysis for phase II was conducted, the pool of participants had grown to 54 participants in total, from which 39 (25 males and 14 females) participants met the inclusion criteria of phase II (described in Section 6.1.2). Participants' ages ranged between 27 and 65 (mean= 33.7, st. dev.= 7.04).

6.1.2 Inclusion Criteria

The inclusion criteria of phase II are composed of the two initial criteria of phase I. The third criterion of phase I (having a sensical response immediately after a nonsensical one) was not relevant to this analysis.

- **Verbal responses:** VH's questions must have been answered verbally. The participant was not considered if responses were typed using a keyboard.
- **Number of empathic opportunities:** During the interaction, at least one empathic opportunity must have been triggered by the participant.

6.1.3 Segmentation of Videos

The video recordings of clinician interactions with VHs that met the inclusion criteria were segmented to capture key empathic moments. During each interaction, specific empathic

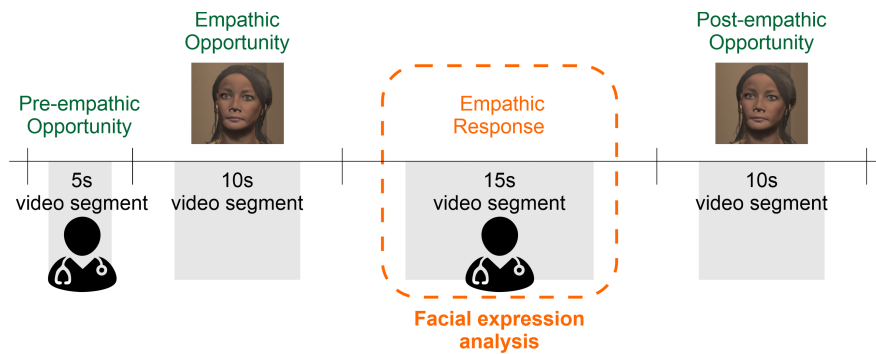


Figure 4. Phase II: Video Segmentation Strategy. Moments of empathic opportunity were segmented into pre-empathic opportunity, empathic opportunity, empathic response, and post-empathic opportunity. We analyzed the facial expressions of clinician-participants using data from the empathic response segments.

opportunities were automatically identified within the transcripts. These moments were categorized into four types:

- **Pre-empathic opportunity:** A clinician prompts an empathic moment, for example, by asking, 'Do you feel like you have lost hope?'
- **Empathic opportunity:** The VH expresses an emotional state, such as, 'I feel like I can't think straight anymore.'
- **Empathic response:** The clinician acknowledges the VH's emotional expression with a response like, 'You must have felt ashamed or humiliated.'
- **Post-empathic opportunity:** The interaction continues with the clinician asking further questions, such as, 'Do you ever feel like you are a burden on others?'

These empathic response segments, which capture the clinician's reactions, were later analyzed to identify and classify the clinicians' basic facial emotions (See Figure 4).

6.2 Metrics

The metrics utilized in phase II were the proportion of VHs' nonsensical responses over the total number of VHs' responses (during each clinician-VH interaction), the average percentage of time clinicians demonstrated each individual emotion (during empathic responses), and clinicians' average empathy scores (calculated based on clinicians' empathic responses, as measured by the ECCS scale). We next describe each measure in detail.

6.2.1 Basic facial emotions

We analyzed clinicians' facial expressions using the empathic response segments described in section 6.1.3. Noldus FaceReader (see Section 4.2) classified facial expressions into basic facial emotions, and provided a percentage of time each emotion was identified per video segment. For each clinician-VH interaction, we averaged the percentage of time clinicians demonstrated each of the seven basic facial emotions and used statistical methods to analyze the data.

6.2.2 Nonsensical responses

VH's nonsensical responses were those that had no meaning, direction, or purpose in relation to the questions posed by

the participants (see Table 1). To quantify the nonsensical responses, the transcripts of the interactions between clinicians and VHs were manually coded by a research staff team. Each instance of sensical and nonsensical response was counted and the percentage of nonsensical responses over the total number of responses was calculated during the interactions with Cynthia and Bernie for each clinician.

6.2.3 ECCS scale

The ECCS scale categorizes empathic responses into three distinct levels [Bylund and Makoul, 2005]:

- **Level 1:** The empathic response either ignores or only implicitly acknowledges the central issue presented in the empathic opportunity, indicating a minimal engagement with the patient's emotional needs.
- **Level 2:** The empathic response explicitly recognizes the central issue and may include additional questions to further explore the patient's concerns, demonstrating a moderate level of empathy and understanding.
- **Level 3:** The empathic response not only acknowledges the central issue but also validates the feelings expressed by the patient, reflecting a high level of empathy and emotional attunement.

Each clinician was assigned an ECCS score based on the level of empathy demonstrated in their responses to empathic opportunities. These scores provided a quantitative measure of the clinician's ability to engage empathetically with patients. Clinicians' responses were systematically coded according to the ECCS levels, and an average empathy score was calculated for their interactions with Virtual Humans Cynthia and Bernie. This average score served as an indicator of the overall quality of the clinician's empathic communication throughout the interactions, offering valuable insights into their ability to connect with their patients emotionally.

6.2.4 Inter-Rater Reliability

To establish a reliable ground truth classification for sensical and nonsensical responses and the ECCS scale, a team of four psychology research assistants, who were already experienced in coding transcripts for the project, underwent additional training led by one of the research team's experts. These assistants served as raters for identifying sensical and nonsensical responses and classifying empathic responses

Figure 5. Multilevel Regressions Examining Interactions between Facial Expressions and Nonsensical Responses in Predicting Average ECCS Scores.

Predictor	Fixed Effects				Random Effects	
	B	SE	p	95% CI	τ_{00}	σ^2
(Intercept)	2.13	.12	<.001	1.88, 2.37		
Time	.10	.13	.476	-.18, .36		
Neutral	-.08	.08	.325	-.25, .10	.17	.23
NS	-.02	.09	.861	-.20, .17		
Neutral x NS	-.04	.09	.635	-.22, .13		
(Intercept)	2.16	.12	<.001	1.93, 2.39		
Time	.03	.12	.827	-.22, .28		
Surprised	.1	.08	.248	-.07, .27	.17	.22
NS	.02	.09	.854	-.17, .20		
Surprised x NS	-.05	.07	.467	-.19, .09		
(Intercept)	2.11	.11	<.001	1.89, 2.33		
Time	.04	.11	.717	-.19, .27		
Angry	-.2	.08	.017	-.35, -.04	.17	.19
NS	.01	.09	.921	-.16, .18		
Angry x NS	.2	.08	.012	.05, .35		
(Intercept)	2.16	.12	<.001	1.92, 2.39		
Time	.07	.13	.584	-.19, .33		
Happy	.004	.08	.959	-.15, .15	.16	.24
NS	.03	.09	.789	-.17, .22		
Happy x NS	.2	.21	.352	-.24, .63		
(Intercept)	2.14	.12	<.001	1.91, 2.38		
Time	.08	.13	.548	-.19, .34		
Sadness	.15	.12	.23	-.11, .39	.16	.23
NS	-.003	.09	.975	-.18, .18		
Sadness x NS	-.07	.17	.689	-.40, .28		
(Intercept)	2.16	.12	<.001	1.91, 2.40		
Time	.04	.13	.75	-.23, .31		
Disgust	-.12	.1	.225	-.31, .07	.14	.24
NS	.01	.1	.929	-.19, .21		
Disgust x NS	.02	.19	.922	-.37, .40		
(Intercept)	2.15	.12	<.001	1.92, 2.39		
Time	.06	.13	.659	-.21, .33		
Fear	.05	.08	.529	-.11, .22	.15	.24
NS	-.02	.09	.825	-.21, .17		
Fear x NS	.03	.05	.492	-.06, .13		

Note: NS = Nonsensical Responses; CI = Confidence Interval. All predictors were standardized prior to inclusion in models.

into the ECCS scale. Raters manually coded the transcripts of interactions between clinicians and virtual humans. Interrater reliability, a key metric to assess the consistency of the ratings, was confirmed during the training phase, achieving a score greater than 0.8. This process quantifies the level of agreement among two or more independent coders assessing the same set of subjects [Hallgren, 2012]. In instances where raters disagreed, differences were resolved through group discussions. If consensus could not be reached, the final gold labels were determined by a majority vote.

6.3 Data Analysis

A series of linear mixed models [Bryk and Raudenbush, 1992] was conducted to examine whether different proportions of nonsensical responses moderated the relationship between the basic facial emotions (i.e., neutral, surprise, anger, happiness, sadness, disgust, fear) exhibited by clinicians and their average empathy as measured by the ECCS scores. Because each clinician completed two interview sessions (one with each VH), session was nested within clinicians as a random intercept to account for the non-independence of observations. Simple slopes of all interaction effects were examined at low (i.e., -1 Standard Deviation (SD)), moderate (i.e., mean), and high (i.e., +1 SD) levels of nonsensical re-

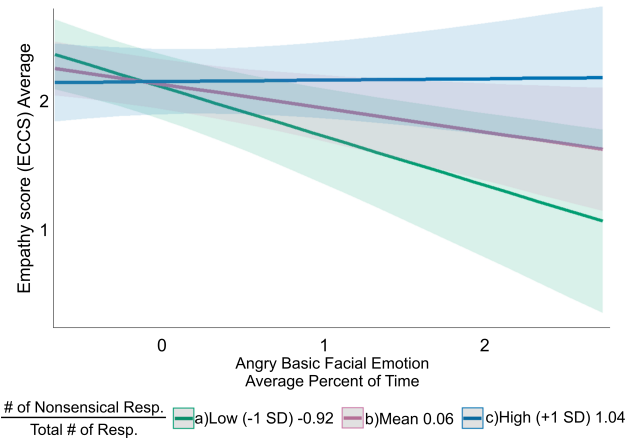


Figure 6. Simple Slopes of Interaction Effect between Angry Basic Facial Emotion and Nonsensical Responses in Predicting Average ECCS Scores.

sponses, consistent with established guidelines [Cohen et al., 2014]. Standardized coefficients with associated 95% confidence intervals are presented to facilitate the interpretation of effect size. All analyses were conducted in R using the lme4 [Bates et al., 2014] and lmerTest [Kuznetsova et al., 2017] packages. Where necessary, we abbreviate Standard Deviation (SD), Standard Error (SE), and the regression coefficient estimate (B).

6.4 Results

Models examining interactions between each facial emotion and nonsensical responses in predicting average ECCS scores were estimated. Detailed statistics for each model are presented in Figure 5. Simple slope analyses at low (-1 SD), mean, and high (+1 SD) levels of nonsensical responses are presented in Figure 6. Next, we summarize the findings for each of the seven basic emotions.

- **Neutral.** The main and interaction effects between happy facial emotion and nonsensical responses were not significantly associated with average ECCS scores.
- **Surprise.** There were no significant main or interaction effects between surprised facial emotions and nonsensical responses in predicting average ECCS scores.
- **Anger.** There was a significant interaction between angry facial emotions and nonsensical responses in predicting average ECCS scores. Specifically, the negative relationship between angry facial emotions and average ECCS scores was:
 1. **Lower levels:** Statistically significant and stronger at lower levels (-1 SD) of nonsensical responses ($B = -.38, SE = .12, p = .002$);
 2. **Mean levels:** Statistically significant at mean levels of nonsensical responses ($B = -.18, SE = .08, p = .022$), and
 3. **High levels:** Non-significant at high levels (+1 SD) of nonsensical responses ($B = .01, SE = .10, p = .916$).

Figure 6 demonstrates the form of the interaction effect.

- **Happy.** The main and interaction effects between happy facial emotions and nonsensical responses were not significantly associated with average ECCS scores.

- **Sad.** The main and interaction effects between sad facial emotions and nonsensical responses were not significantly associated with average ECCS scores.
- **Disgust.** The main and interaction effects between disgusted facial emotions and nonsensical responses were not significantly associated with average ECCS scores.
- **Scared.** There were no significant main or interaction effects between scared facial emotions and nonsensical responses in predicting average ECCS scores.

6.5 Discussion

The results show a significant interaction between the angry basic emotion and nonsensical responses in predicting average ECCS scores (See Figures 5 and 6). Specifically, there was a significant negative relationship between angry facial expressions and average ECCS scores at low levels ($B = -.38$, $SE = .12$, $p = .002$) of nonsensical responses and at mean levels ($B = -.18$, $SE = .08$, $p = .022$) of nonsensical responses. However, no significant relationship was found at high levels ($B = .01$, $SE = .10$, $p = .916$) of nonsensical responses. Additionally, no significant relationships were observed for the other basic emotions analyzed.

These results help us further understand the relationship between nonsensical responses, basic emotions, and empathy scores. The results of phase I suggest that nonsensical responses from VHs can increase the percentage of time clinicians show NFABs while interacting with VHs. NFABs were defined by grouping the anger, sadness, scare, and disgust emotions. The results of phase I led us to hypothesize that NFABs could potentially hinder VHs' effectiveness during training and therefore negatively impact users' performance. The results of Phase II support that hypothesis. Specifically, the results of phase II suggest that there is a distinct role of anger among the other NFABs. This was a surprising result since we expected that several negative facial behaviors would be differently affected by different levels of nonsensical responses.

6.5.1 Implications of Low and Mean Levels of Nonsensical Responses

The results suggest a significant negative relationship between anger and empathy scores for lower and mean levels of nonsensical responses (See Figure 6), showing that at those levels of nonsensical responses clinicians' growing anger significantly impacted training performance (as measured by clinicians' empathy scores) in a negative way. These findings are consistent with prior work that found that increased anger in learning environments can negatively impact learners' performance [Assor et al., 2005].

Based on the results associated with exposure to low and mean levels of nonsensical responses, we suggest that VH researchers and developers must control for nonsensical responses and communication issues in general toward increasing participants' performance during training. Perhaps more importantly, a major implication of these results is to highlight the importance of controlling for participants' emotions during the VH-based simulation, specifically anger, toward maximizing training effectiveness.

6.5.2 Implications of High Levels of Nonsensical Responses

Surprisingly, at high levels of nonsensical responses, we did not observe a significant relationship between the anger emotion and empathy scores ($B = .01$, $SE = .10$, $p = .916$) (See Figure 6). This means that no relationship can be drawn between the angry facial expressions elicited and the ECCS scores obtained by clinicians when there were high levels of nonsensical responses. Put simply, clinicians exposed to high levels of nonsensical responses obtained scores ranging from high to low without a significant relationship with the nonsensical responses and emotions elicited.

While we can't derive any conclusions regarding high levels of nonsensical responses, we suspect that, at all levels, nonsensical responses may have *distracted* clinicians and critically disrupted their *engagement* with the VH simulation. The results obtained at low and mean levels of nonsensical responses are in alignment with Wang et al. [2013], which found that distractions caused by VHs' communication errors negatively influenced participants' performance. The results obtained at low and mean levels of nonsensical responses are also in alignment with prior work that has observed that a lack of engagement may hinder the effectiveness of the simulation [Assor et al., 2005]. Additionally, nonsensical responses may also have negatively impacted clinicians' willingness to "suspend disbelief" during the VH simulation, further contributing to a loss of engagement. These hypotheses may help explain the results obtained at low and mean levels of nonsensical responses. We next analyze each of these factors: engagement and suspension of disbelief.

6.5.3 Nonsensical Responses and Engagement

To support the hypothesis that clinicians' engagement may have decreased due to the increasing levels of nonsensical responses—potentially exacerbated by the anger these responses elicited—we analyzed two proxy measures: clinicians' perceptions of the *perceived value* and *ease of use* of the VH interaction as a training tool.

Previous studies have shown that higher levels of user engagement are positively correlated with both perceived value and ease of use [Kim et al., 2013; Davis and Wiedenbeck, 2001]. These relationships informed our decision to use these two measures as proxies for engagement, given that the original data collection protocol did not consider collecting data on clinicians' engagement with the VH interaction.

After interacting with the second virtual patient, clinicians completed the "Virtual Human Interaction Satisfaction Survey" [Foster et al., 2016], which included several questions assessing their experience (See Section 4.3). For this analysis, we focused on two specific questions: their perceptions of the value of VH interaction as a training tool and the ease of use of the VH platform. Both questions were rated on a 5-point Likert scale, ranging from poor to excellent.

We conducted an exploratory follow-up analysis to examine how variations in specific input variables, such as the level of nonsensical responses, influenced the outcomes. This analysis utilized Pearson's correlation coefficient to assess the relationships between these variables. This approach

revealed a negative relationship between the frequency of nonsensical responses from the VHS and clinicians' perceptions of ease of use ($r = -.30, p = .007$). A similar negative correlation was found between nonsensical responses and the perceived value of the VH interaction as a training tool ($r = -.25, p = .022$). While these findings are preliminary, they suggest that as the frequency of nonsensical responses increased, clinicians rated both the ease of use and the training value of the VH platform lower. These results highlight the diverse negative implications of increasing nonsensical responses, which can lead to distraction and disengagement in VH-based training simulations, ultimately undermining their effectiveness. Additionally, these findings underscore the need for further research to directly examine engagement, particularly how communication issues like nonsensical responses impact clinician engagement. This deeper exploration will be crucial to fully understanding and enhancing the overall success of training programs.

6.5.4 Nonsensical Responses and Suspension of Disbelief

In training simulations involving VHS, participants must accept the otherwise unrealistic aspects of the clinical simulation for the training simulation to be effective [Muckler, 2017]. Applying to the context of this study, clinicians must be willing to "suspend disbelief" that VHS are "real" and deserving of empathy. Factors that contribute to participants' ability to suspend disbelief include fidelity, emotional buy-in, and the fiction contract [Muckler, 2017]. We suggest that increasing levels of nonsensical responses can negatively impact each of these factors.

It is possible that nonsensical responses may have impacted participants' perceived VH fidelity (real humans would not respond with such nonsensical responses), their emotional buy-in (participants may feel less inclined to feel empathy toward VHS that give nonsensical responses than real humans), and they may have felt that the fiction contract had been broken (VHS were supposed to behave like real humans in the simulation.) Additionally, others have observed that we are willing to suspend disbelief to gain pleasure [Holland, 1967]. However, nonsensical responses can be argued to promote experiences that are the opposite of pleasurable. Indeed, in a post-interaction survey (responded after interacting with the second VH,) clinicians from all levels of nonsensical responses mentioned frustration and difficulty in believing that Cynthia and Bernie were human. One clinician wrote, "I felt discouraged and frustrated since the VHS struggled to understand what I was trying to say or ask. This does not come close to communication and assessment with an actual human." Another clinician mentioned, "It was hard to pretend it was a real patient." These comments suggest an inability of clinicians exposed to nonsensical responses to form a significant relationship with VHS and further support the idea that difficulties in communication negatively impacted clinicians' ability to suspend disbelief, ultimately contributing to a loss of engagement.

6.5.5 Nonsensical Responses and Negative Emotions

These results may have relevant methodological implications for those researching and utilizing systems that employ VHS in training simulations, particularly regarding technological failures and the emotions users may experience. Technological limitations that impact communication (such as those that cause nonsensical responses) may hinder clinicians' empathy skills acquisition and emotional training. Moreover, failure to control for clinicians' emotions (in particular anger) during real-world or virtual clinician-patient interaction can undermine verbal empathic communication. During training, anger emotions may limit clinicians' ability to acquire skills. During real-world interactions with patients, clinicians demonstrating anger may be perceived as lacking empathy. In sum, controlling for clinicians' emotions, specifically for anger, and controlling for technologically motivated communication failures may be of major relevance in VH training simulations.

6.5.6 Nonsensical Responses and AI-based Conversational Models

Our study, which examines the use of virtual human patients to train clinicians in empathic communication, remains highly relevant even as Artificial Intelligence (AI) conversational models advance. While modern AI systems have improved in generating human-like interactions, they still produce nonsensical responses, particularly in the form of hallucinations. These errors can disrupt conversations and negatively impact training outcomes, much like the nonsensical responses observed in our VH interactions.

In the broader context of AI, our findings underscore a critical point: despite their advancements, AI conversational models can still generate responses that disrupt sensitive interactions, such as those involved in clinical training. Our study provides valuable insights into how these disruptions affect performance and emphasizes the need for careful consideration of these issues. By addressing these challenges, we can improve AI systems in training and other applications where empathic communication is essential.

7 Overall Conclusion

This study presents our analysis of the effects of communication issues during clinician-VH interactions in a simulation that aimed to train clinicians on empathic communication skills. During training, clinicians received empathy scores as they interacted with two VHS portraying suicidal ideation, who at times provided nonsensical responses. We analyzed clinicians' empathy scores and basic facial emotions as they were exposed to varying levels of nonsensical responses. In phase I, we grouped emotions as Positive (PFABs) and Negative (NFABs). The results suggest that, after a nonsensical response, participants demonstrate a statistically significant increase in NFABs when compared to after a sensible response. This key result of phase I motivated phase II of this study.

Phase II analyzed whether the proportion of nonsensical responses during clinicians' interactions with VHS moderates the relationship between basic emotions and the average

empathy scores obtained by clinicians during training. The results show a statistically significant interaction between levels of nonsensical responses and emotions (as elicited by facial expression analysis) in predicting average empathy scores. The relationship between angry facial expressions and empathy scores was significant at low and mean, but non-significant at high levels of nonsensical responses. We suggest that nonsensical responses may impact users' engagement with the VH simulation and their ability to suspend disbelief toward VH systems. These factors may hinder the simulation's effectiveness as a training tool, leading to the observed impact on empathy scores. More broadly, the results suggest that the anger emotion identified by facial analysis and promoted by communication failures may hinder aspects of the clinician-patient verbal communication, which is critical of any human interaction.

Our study, which examines the use of virtual human (VH) patients to train clinicians in empathic communication, remains highly relevant even as AI-based conversational models advance. While modern AI systems have improved in generating human-like interactions, they still produce nonsensical responses, particularly in the form of hallucinations. These errors can disrupt conversations and negatively impact training outcomes, much like the nonsensical responses observed in our VH interactions.

Our research found that nonsensical VH responses led to increased negative emotions in clinicians, such as anger and sadness, which in turn hindered their ability to develop empathy skills. Notably, these negative impacts were most pronounced when the proportion of nonsensical responses was low to moderate. This highlights the importance of managing such issues, even in sophisticated AI systems.

Overall, this study has demonstrated a viable approach to identifying clinicians' emotions during training and the impact of technologically motivated communication failures on empathic training effectiveness. We hope this work can motivate the research community to broadly investigate and consider technological and communication issues that impact users' emotions and their effects on training outcomes, particularly in simulations involving VHS and natural language processing systems in the context of difficult conversations.

7.1 Future Work and Limitations

Beyond VH training simulations, this work may be relevant to broader contexts. For instance, in real-world scenarios involving telehealth therapy conducted over video conferencing, a poor connection will likely represent a technological issue that may elicit negative emotions (such as anger) and impede clinicians' ability to empathize with patients. Future work may show that controlling for users' emotions and technological issues may increase the effectiveness in such scenarios too. Future work should focus on characterizing the threshold at which participants' performance is impacted by such technological failures. Put simply, others should attempt to determine the minimum tolerable percentage of nonsensical responses that participants can be exposed to without critically hindering performance during training. Additionally, regarding the anger emotion, future work could explore strategies to support anger management during training. Ex-

pressing and controlling one's anger have been shown to differently affect learning performance [Boekaerts, 1994] and could be coupled with efforts that aim to mitigate technological issues. For instance, future work may explore the impact of treatments that aim to elicit positive emotions in participants. It is possible that positive emotions could counterbalance negative ones and potentially enhance training effectiveness, leading to increased user performance. Overall, such efforts are relevant since they can potentially demonstrate the usefulness of less-than-perfect systems, with important implications for the time and cost related to developing and maintaining solutions that employ complex VHS and dialogue systems. It may be especially relevant to systems that have been maintained and adapted over the course of multiple iterations (perhaps for years) by different groups of researchers and specialists, a scenario that may become more common as the number of solutions in use grows and VH technology matures.

The results presented in this work are based on facial expression analysis. A limitation of this approach is that users who preferred to communicate with VHS using a keyboard often looked down while typing and could not be considered in the study. Noldus FaceReader requires that the users' faces be fully captured to be analyzed. Toward overcoming this limitation, future work should investigate how nonsensical responses affect the measurements obtained with other technologies such as galvanic skin response and heartbeat rate sensors which do not require users' facial expressions to be captured. The results obtained in this work were observed as nonsensical responses emerged during the clinician-VH interactions, for example, due to issues in the conversational model or during the speech-to-text conversion. This means that participants were not exposed to predetermined nonsensical moments. This approach limits our ability to analyze the passage of time and how exposure to nonsensical responses in different moments affects the impact of nonsensical responses on participants' emotions. While the current approach reproduces what may happen in the real world (enhancing external validity), exposing participants to nonsensical responses at predetermined moments may allow for the analysis of how the effects of nonsensical responses change over time. For instance, others have found that drops in reliability after a period of good performance are in general more harmful to users' performance than early failures [Desai et al., 2012, 2013]. Additionally, in this study, each participant may have received different nonsensical responses depending on the questions they posed to the VHS. Exposing participants to the same set (and number) of nonsensical responses may increase the internal validity of the study.

Declarations

Acknowledgements

This is a multiline text of acknowledgments. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Funding

This work was supported by the American Foundation for Suicide Prevention [LSRG-1-050-18] and the National Institutes of Health [NIMH 1R34MH119294-01].

Authors' Contributions

Alexandre Gomes de Siqueira: Conceptualization, Methodology, Software, Writing- Original draft preparation, Investigation. Heng Yao: Conceptualization, Methodology, Software, Writing- Reviewing and Editing. Megan L. Rogers: Conceptualization, Methodology, Investigation, Validation, Writing- Reviewing and Editing. Olivia C. Lawrence, Devon Peterkin, Sifan Zheng, Kathleen Feeney: Software validation, Writing- Reviewing and Editing. Erica D. Musser: Software, Validation, Writing- Reviewing and Editing. Igor Galynker: Conceptualization, methodology, Supervision. Benjamin Lok: Conceptualization, Methodology, Software, Supervision.

Competing interests

The authors AGS, HY, MLR, OCL, DP, KF, SZ, EDM, and IG declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The author Bk declares the following financial interests/personal relationships which may be considered as potential competing interests: Benjamin Lok reports financial support was provided by the National Institutes of Health. Benjamin Lok reports a relationship with ELSEVIER INC that includes: consulting or advisory. Benjamin Lok has the patent COMMUNICATION AND SKILLS TRAINING USING INTERACTIVE VIRTUAL HUMANS licensed to Elsevier.

References

- Assor, A., Kaplan, H., Kanat-Maymon, Y., and Roth, G. (2005). Directly controlling teacher behaviors as predictors of poor motivation and engagement in girls and boys: The role of anger and anxiety. *Learning and instruction*, 15(5):397–413. DOI: 10.1016/j.learninstruc.2005.07.008.
- Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*. DOI: 10.18637/jss.v067.i01.
- Bloch-Elkouby, S., Barzilay, S., Gorman, B. S., Lawrence, O. C., Rogers, M. L., Richards, J., Cohen, L. J., Johnson, B. N., and Galynker, I. (2021). The revised suicide crisis inventory (sci-2): Validation and assessment of prospective suicidal outcomes at one month follow-up. *Journal of affective disorders*, 295:1280–1291. DOI: 10.1016/j.jad.2021.08.048.
- Bloch-Elkouby, S., Gorman, B., Schuck, A., Barzilay, S., Calati, R., Cohen, L. J., Begum, F., and Galynker, I. (2020). The suicide crisis syndrome: A network analysis. *Journal of counseling psychology*, 67(5):595. DOI: 10.1037/cou0000423.
- Boekaerts, M. (1994). Anger in relation to school learning. *Learning and instruction*, 3(4):269–280. DOI: 10.1016/0959-4752(93)90019-V.
- Borish, M., Cordar, A., Foster, A., Kim, T., Murphy, J., and Lok, B. (2014). Utilizing real-time human-assisted virtual humans to increase real-world interaction empathy. *Kansei Engineering Emotion Research (KEER'14)*, 15. DOI: 10.1109/IM.2003.1240281.
- Bryk, A. S. and Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Sage Publications, Inc. DOI: 10.1080/00401706.1994.10485413.
- Bylund, C. L. and Makoul, G. (2002). Empathic communication and gender in the physician–patient encounter. *Patient education and counseling*, 48(3):207–216. DOI: 10.1016/s0738-3991(02)00173-8.
- Bylund, C. L. and Makoul, G. (2005). Examining empathy in medical encounters: an observational study using the empathic communication coding system. *Health communication*, 18(2):123–140. DOI: 10.1207/s15327027hc1802_2.
- Cohen, A. S., Morrison, S. C., and Callaway, D. A. (2013). Computerized facial analysis for understanding constricted/blunted affect: initial feasibility, reliability, and validity data. *Schizophrenia research*, 148(1-3):111–116. DOI: 10.1016/j.schres.2013.05.003.
- Cohen, P., West, S. G., and Aiken, L. S. (2014). *Applied multiple regression/correlation analysis for the behavioral sciences*. Psychology press. DOI: 10.4324/9780203774441.
- Cootes, T. and Taylor, C. (2004). Statistical models of appearance for computer vision, imaging science and biomedical engineering. *University of Manchester*, 17:2011. DOI: 10.1007/3-540-44935-3_47.
- Davis, S. and Wiedenbeck, S. (2001). The mediating effects of intrinsic motivation, ease of use and usefulness perceptions on performance in first-time and subsequent computer users. *Interacting with computers*, 13(5):549–580. DOI: 10.1016/s0953-5438(01)00034-0.
- Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., and Yanco, H. (2013). Impact of robot failures and feedback on real-time trust. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 251–258. IEEE. DOI: 10.1109/hri.2013.6483596.
- Desai, M., Medvedev, M., Vázquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., Steinfeld, A., and Yanco, H. (2012). Effects of changing reliability on trust of robot systems. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 73–80. IEEE. DOI: 10.1145/2157689.2157702.
- DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., et al. (2014). Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068. DOI: 10.1609/aaai.v29i1.9777.
- Dixon, W. J. and Mood, A. M. (1946). The statistical sign test. *Journal of the American Statistical Association*, 41(236):557–566. DOI: 10.1037/e417432004-001.
- Ekman, P. and Cordaro, D. (2011). What is meant by calling emotions basic. *Emotion review*, 3(4):364–370. DOI: 10.1177/1754073911410740.

- Ekman, P., Sorenson, E. R., and Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. *Science*, 164(3875):86–88. DOI: 10.1126/science.164.3875.86.
- Foster, A., Chaudhary, N., Kim, T., Waller, J. L., Wong, J., Borish, M., Cordar, A., Lok, B., and Buckley, P. F. (2016). Using virtual patients to teach empathy: a randomized controlled study to enhance medical students' empathic communication. *Simulation in Healthcare*, 11(3):181–189. DOI: 10.1097/sih.000000000000142.
- Galynker, I. (2017). *The suicidal crisis: Clinical guide to the assessment of imminent suicide risk*. Oxford University Press. DOI: 10.1093/med/9780197582718.002.0001.
- Gomes de Siqueira, A., Yao, H., Bafna, A., Bloch-Elkouby, S., Richards, J., Lloveras, L. B., Feeney, K., Morris, S., Musser, E. D., Lok, B., et al. (2021). Investigating the effects of virtual patients' nonsensical responses on users' facial expressions in mental health training scenarios. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*, pages 1–10. DOI: 10.1145/3489849.3489864.
- Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23. Available at: <https://pmc.ncbi.nlm.nih.gov/articles/PMC3402032/#S20>.
- Holland, N. N. (1967). The "willing suspension of disbelief" revisited. *Centennial Review*, pages 1–23. DOI: 10.5040/9781474218627.ch-001.
- Kim, S. S., Kaplowitz, S., and Johnston, M. V. (2004). The effects of physician empathy on patient satisfaction and compliance. *Evaluation the health professions*, 27(3):237–251. DOI: 10.1177/0163278704267037.
- Kim, Y. H., Kim, D. J., and Wachter, K. (2013). A study of mobile user engagement (moen): Engagement motivations, perceived value, satisfaction, and continued engagement intention. *Decision support systems*, 56:361–370. DOI: 10.1016/j.dss.2013.07.002.
- Kleinsmith, A., Rivera-Gutierrez, D., Finney, G., Cendan, J., and Lok, B. (2015). Understanding empathy training with virtual patients. *Computers in human behavior*, 52:151–158. DOI: 10.1016/j.chb.2015.05.033.
- Krieger, J. L., Neil, J. M., Duke, K. A., Zalake, M. S., Tavasoli, F., Vilaro, M. J., Wilson-Howard, D. S., Chavez, S. Y., Laber, E. B., Davidian, M., et al. (2021). A pilot study examining the efficacy of delivering colorectal cancer screening messages via virtual health assistants. *American journal of preventive medicine*. DOI: 10.1016/j.amepre.2021.01.014.
- Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of statistical software*, 82:1–26. DOI: 10.18637/jss.v082.i13.
- Lewinski, P., den Uyl, T. M., and Butler, C. (2014). Automated facial coding: validation of basic emotions and faces in a face reader. *Journal of Neuroscience, Psychology, and Economics*, 7(4):227. DOI: 10.1037/npe0000028.supp.
- Lucas, G. M., Boberg, J., Traum, D., Artstein, R., Gratch, J., Gainer, A., Johnson, E., Leuski, A., and Nakano, M. (2018). Culture, errors, and rapport-building dialogue in social agents. In *Proceedings of the 18th International Conference on intelligent virtual agents*, pages 51–58. DOI: 10.1145/3267851.3267887.
- Muckler, V. C. (2017). Exploring suspension of disbelief during simulation-based learning. *Clinical Simulation in Nursing*, 13(1):3–9. DOI: 10.1016/j.ecns.2016.09.004.
- Nass, C., Steuer, J., and Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 72–78. DOI: 10.1016/b978-155860643-2/50007-x.
- Pinto, M. D., Greenblatt, A. M., Hickman, R. L., Rice, H. M., Thomas, T. L., and Clochesy, J. M. (2015). Assessing the critical parameters of esmart-mh: A promising avatar-based digital therapeutic intervention to reduce depressive symptoms. *Perspectives in psychiatric care*, 52(3):157–168. DOI: 10.1111/ppc.12112.
- Rizzo, A., Reger, G., Perlman, K., Rothbaum, B., Difede, J., McLay, R., Graap, K., Gahm, G., Johnston, S., Deal, R., et al. (2011). Virtual reality posttraumatic stress disorder (ptsd) exposure therapy results with active duty oif/oef service members. DOI: 10.1515/ijdh.2011.060.
- Ross, J. and Watling, C. (2017). Use of empathy in psychiatric practice: constructivist grounded theory study. *BJPsych open*, 3(1):26–33. DOI: 10.1192/bjpo.bp.116.004242.
- Rossen, B. and Lok, B. (2012). A crowdsourcing method to develop virtual human conversational agents. *International Journal of Human-Computer Studies*, 70(4):301–319. DOI: 10.1016/j.ijhcs.2011.11.004.
- Schuck, A., Calati, R., Barzilay, S., Bloch-Elkouby, S., and Galynker, I. (2019). Suicide crisis syndrome: A review of supporting evidence for a new suicide-specific diagnosis. *Behavioral sciences the law*, 37(3):223–239. DOI: 10.1002/bsl.2397.
- Skarbez, R., Kotranza, A., Brooks, F. P., Lok, B., and Whittton, M. C. (2011). An initial exploration of conversational errors as a novel method for evaluating virtual human experiences. In *2011 IEEE Virtual Reality Conference*, pages 243–244. IEEE. DOI: 10.1109/vr.2011.5759489.
- Stevens, A., Hernandez, J., Johnsen, K., Dickerson, R., Rajj, A., Harrison, C., DiPietro, M., Allen, B., Ferdig, R., Foti, S., et al. (2006). The use of virtual patients to teach medical students history taking and communication skills. *The American Journal of Surgery*, 191(6):806–811. DOI: 10.5005/jp/books/12827_18.
- Stuart, J., Akinola, I., Guido-Sanz, F., Anderson, M., Diaz, D., Welch, G., and Lok, B. (2020). Applying stress management techniques in augmented reality: Stress induction and reduction in healthcare providers during virtual triage simulation. In *2020 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops (VRW)*, pages 171–172. IEEE. DOI: 10.1109/vrw50115.2020.00037.
- Stuart, J., Aul, K., Bumbach, M. D., Stephen, A., de Siqueira, A. G., and Lok, B. (2022). The effect of virtual humans making verbal communication mistakes on learners' perspectives of their credibility, reliability, and trustworthiness. In *2022 IEEE Conference on Virtual Reality*

- and *3D User Interfaces (VR)*, pages 455–463. IEEE. DOI: 10.1109/vr51125.2022.00065.
- Wang, Y., Khooshabeh, P., and Gratch, J. (2013). Looking real and making mistakes. In *International Workshop on Intelligent Virtual Agents*, pages 339–348. Springer. DOI: 10.1007/978-3-642-40415-3_30.
- Wind, L. A., Van Dalen, J., Muijtjens, A. M., and Rethans, J.-J. (2004). Assessing simulated patients in an educational setting: the masp (maastricht assessment of simulated patients). *Medical Education*, 38(1):39–44. DOI: 10.1111/j.1365-2923.2004.01686.x.
- Yao, H., de Siqueira, A. G., Foster, A., Galynker, I., and Lok, B. (2020). Toward automated evaluation of empathetic responses in virtual human interaction systems for mental health scenarios. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8. DOI: 10.1145/3383652.3423916.