AMERICAN PSYCHOLOGICAL ASSOCIATION

Technology, Mind, and Behavior

ISSN: 2689-0208

https://doi.org/10.1037/tmb0000164

SINGLE-STUDY PAPER

Trustworthy Enough? Examining Trustworthiness Assessments of Large Language Model-Based Medical Agents

Nadine Schlicker[1], Fabian Lechner[1], Katja Wehrle[2], Berit Greulich[3], Martin C. Hirsch[1], and Markus Langer[4]

[1] Department of Medicine, Philipps-Universität Marburg
[2] Department of Psychology, Justus-Liebig-Universität Gießen
[3] Department of Medicine, Universität des Saarlandes
[4] Department of Psychology, Albert–Ludwigs–Universität Freiburg

This research advances trust theory by examining factors shaping the development of a trustor's perceived trustworthiness in the context of real-world interactions with a large language model-driven virtual doctor (VD). Employing a qualitative approach to elaborate the trustworthiness assessment model, we conducted 51 interviews with 65 participants. Our findings reveal a heterogeneity in the trustworthiness perceptions of and reported trust in VDs, ranging from a complete absence to a complete presence of trust, with many participants expressing conditional trust. The key factors contributing to this heterogeneity were participants' benchmarks for trustworthiness, naïve theories, risk–benefit assessments, individual standards, and strategies for cue detection and utilization in assessing the trustworthiness of the VD. Our findings also highlight the crucial influence of third-party involvement in artificial intelligence system development and testing on trustworthiness assessments. These insights underscore the trustworthiness assessment model's utility in understanding trust development processes.

Keywords: large language models, trust in artificial intelligence, trustworthiness assessment, artificial intelligence in medicine

Trust shapes the dynamics of patient–physician relationships and is a predictor of treatment success (for an overview, see Hall et al., 2001). Seeking care in states of high vulnerability, be they on the account of one's compromised health or the exposure of sensitive personal information, requires individuals to trust in health care providers—irrespective of whether they are human or automated systems (Juravle et al., 2020; Pearson & Raeke, 2000). As such, trust is crucial when it comes to medical artificial intelligence (AI).

Within medical AI, large language models (LLMs) are expected to significantly impact health care (Thirunavukarasu et al., 2023). The present research has reported the performance of LLMs to be comparable or even above human levels. For instance, LLMs passed medical examination tests (Nori et al., 2023; Singhal et al., 2023) and they can perform better than humans when evaluating selected patient cases (Goh et al., 2024). While LLMs are currently rare in professional medical applications, they are already available to and used by lay individuals for answering medical questions (Choudhury & Shamszare, 2023), such as the customGPT "Virtual Doctor" (https://chat.openai.com/gpts). There is a hope that LLMs that provide medical decision support (hereinafter referred to as *virtual doctors [VDs]*) will lead to more cost-effective patient care and less crowded care units (Zhang et al., 2023). Moreover, VDs are considered beneficial by different stakeholders within the health care system and are, as such, a likely tool for patients seeking medical

Action Editor: Danielle S. McNamara was the action editor for this article.

ORCID iD: Nadine Schlicker https://orcid.org/0000-0003-0368-3081.

Change of Affiliation: Nadine Schlicker is now at Institut für KI in der Medizin, Philipps-Universität Marburg.

Funding: This work was partially funded by the Deutsche Forschungsgemeinschaft Grant 389792660 (awarded to Markus Langer) as part of Transregional Collaborative Research Centre 248, Center for Perspicuous Computing. Open access funding was provided by the Open Access Publishing Fund of Philipps-Universität Marburg.

Disclosures: The authors have no conflicts of interest to disclose.

Data Availability: The data supporting the findings of this study are available within the article. Additional data can be accessed on the Open Science Framework repository at https://osf.io/856hq/?view_only=b07f91d14c06435cb87be2eb1e121f20. Further data may be made available upon reasonable request to the corresponding author. Full transcripts and participant-related materials will not be shared due to privacy and confidentiality considerations. There is no prior use of the data.

Contact Information: Correspondence concerning this article should be addressed to Nadine Schlicker, Institut für KI in der Medizin, Philipps-Universität Marburg, Baldingerstraße, 35043 Marburg, Germany. Email: nadine.schlicker@uni-marburg.de

advice in the near future (e.g., Lechner et al., 2023; Loh et al., 2024; Sandmann et al., 2024; Yang et al., 2024).

Although LLMs demonstrate unprecedented performance on certain tasks, they are likely to remain imperfect for the foreseeable future (Ghosh & Joshi, 2020; P. Lee et al., 2023). These imperfections include the production of inaccurate outputs, sometimes fabricating false information (i.e., hallucinating; Thirunavukarasu et al., 2023), as well as a lack of transparency and clinical validation (Gilbert et al., 2023). Consequently, it is important that potential users of VDs (or similar AI-based technologies) critically evaluate the associated risks and potentials to foster well-grounded trust and ensure optimal reliance (de Visser et al., 2014; McLeod, 2023; Wischnewski et al., 2023). While trust is an important factor in the adoption of VDs (van Bussel et al., 2022; Wutz et al., 2023), both excessive and insufficient trust can have detrimental effects on patients' well-being; inaccurate outputs might go undetected and the full potential of AI-based technology might be underexploited (Choudhury & Shamszare, 2023; Dietvorst et al., 2015). The development of appropriate trust and reliance largely depends on a trustor's *assessment of a trustee's trustworthiness* (Mayer et al., 1995; Schlicker et al., 2025). Whereas many theories emphasize the importance of perceived trustworthiness in the trust development process (J. D. Lee & See, 2004; Mayer et al., 1995) and acknowledge that *actual* trustworthiness (of the trustee) and trustor's *perceived* trustworthiness (of the trustee) may not always align (de Visser et al., 2020), the exact mechanisms of how trustor's reach their perceived trustworthiness still need specification.

The present study seeks to close this gap by examining how trustors develop their trustworthiness perceptions through the lens of the trustworthiness assessment model (TrAM; Schlicker et al., 2025; Schlicker & Langer, 2021). While the TrAM provides a conceptual framework for understanding how individuals assess the trustworthiness of a system, its propositions have yet to be empirically tested. To address this, alongside responding to recent calls for qualitative research on human interaction with (medical) AI to advance trust theory (Okpanum et al., 2024; Qualitative Health Research, 2024; see van Bussel et al., 2022; Zhan et al., 2024, for exceptions for qualitative studies), we conducted a qualitative study. In this study, participants interacted in a field setting with an LLM-based VD that mimicked a natural, humanlike dialogue in the context of medicine. Following their interaction, participants were interviewed to explore their trustworthiness assessments.

The present study offers three main contributions. First, we advance trust theory by empirically testing and specifying the propositions of the TrAM (Schlicker et al., 2025) and by deepening the understanding of the TrAM's relation to established trust models (Hoff & Bashir, 2015; J. D. Lee & See, 2004; Mayer et al., 1995; see Fisher & Aguinis, 2017). These insights specifically shed light on the development of users' perceived trustworthiness. Second, we contribute to research on trust in medical AI by studying a context in which individuals interact with an LLM. Unlike vignette studies that are still the most commonly used research design in this area (e.g., Jin & Eastin, 2024; Juravle et al., 2020; Longoni et al., 2019; Promberger & Baron, 2006; Riedl et al., 2024; van Bussel et al., 2022), our study offers insights into concrete trustworthiness assessments of an actual AI-based system. This approach addresses the need for more realistic research designs that enhance the ecological validity of findings (Lew et al., 2011; van Berkel et al., 2020). Third, we provide initial insights into how patients perceive current LLM-based VDs and the factors influencing these perceptions (see Kollerup et al., 2024, for another recent study on LLM-based VDs), advancing research on earlier forms of conversational agents in medicine (Milne-Ives et al., 2020; Wutz et al., 2023), and enabling the derivation of practical design implications for future development. This is important because VDs are likely to become a part of routine care in the future (Yang et al., 2024; Zhang et al., 2023), and most studies in this area were conducted before the public availability of LLMs (see Choudhury & Shamszare, 2023; Sun et al., 2024, for exceptions).

Theoretical Background

In the following, we summarize the current state of trust (in automation and AI) research based on the seminal work of Mayer et al. (1995) and the recent addition of the TrAM (Schlicker et al., 2025). Figure 1 shows a synthesized model that combines the organizational trust model (Mayer et al., 1995; bottom model within dotted frame on gray background)—which describes the process from a trustor's perceived trustworthiness to trusting behavior—and the TrAM—which focuses on the transition from actual trustworthiness to perceived trustworthiness (Schlicker et al., 2025; top model within gray pointed frame). We call the combination of those theoretical models the *trust development process*, the process that builds the theoretical basis for the current research and qualitative analysis.

Trust (in Automation and AI)

Mayer et al. (1995, p. 712) defined 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.

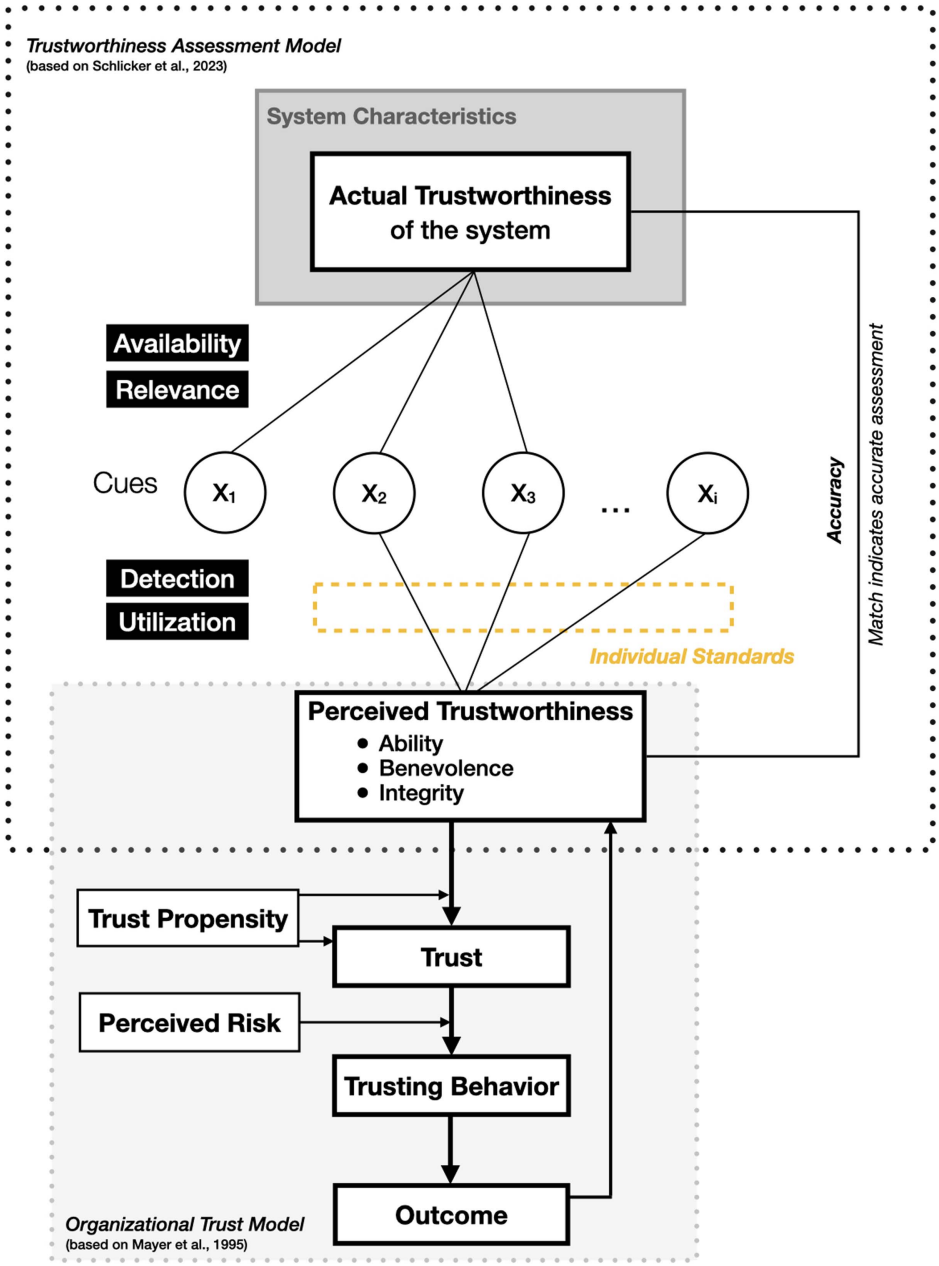
Here, the trustor refers to the person who potentially trusts and the trustee describes the agent (e.g., human or automated system) who is potentially trusted.

According to Mayer et al. (1995), trust depends on the trustor's propensity to trust and their perceived trustworthiness of the trustee (see bottom part of Figure 1). Propensity to trust refers to a trustor's stable disposition to trust other agents, whether they are humans (Mayer et al., 1995) or automated systems (Jessup et al., 2019), regardless of specific contexts or trustees. Perceived trustworthiness, as defined by Mayer et al. (1995), comprises three factors: ability, integrity, and benevolence, which together form the basis of how a trustor assesses a trustee.[1] Trust precedes trusting behavior (Mayer et al., 1995). However, trusting behavior, that is, the act of actually trusting an agent, does not always follow from trust (Bustamante, 2009; Satterfield et al., 2017). Mayer et al. (1995) suggested that the transition from trust to trusting behavior is moderated by the risk assessment in specific situations. That is, trustors weigh the perceived stakes (i.e., the gains and losses of trusting), which influence

[1] We acknowledge that there is a debate about the extent to which the trustworthiness facets proposed by Mayer et al. (1995) apply to the trustworthiness of AI, how many facets of trustworthiness there are, and how they should be called (Glikson & Woolley, 2020; J. D. Lee & See, 2004; Toreini et al., 2020). Nevertheless, there is consensus that trustworthiness is a multifaceted construct.

Figure 1

The Trust Development Process



Note. It shows a synthesized model that combines the organizational trust model (Mayer et al., 1995; bottom model within dotted frame on gray background), which describes the process from a trustor's perceived trustworthiness to trusting behavior, and the trustworthiness assessment model (TrAM; Schlicker et al., 2025; top model within gray dotted frame). The TrAM proposes that the actual trustworthiness of a system is not directly observable and thus inferred by trustors via observable cues (circles). Furthermore, the TrAM proposes that the actual trustworthiness of a system depends on the trustor's individual standards (yellow dashed frame), that is, the norms and values that define a trustworthy system from the trustor's perspective. The actual trustworthiness reflects the "true" value of trustworthiness of the system, that is, how closely the system characteristics align with a trustor's individual standards, assuming a complete assessment of the system characteristics could be conducted. While theoretically possible, achieving this complete assessment is not feasible in practice—for instance, due to the impossibility of testing a system on the entire target population. The perceived trustworthiness reflects how much trustors think that the system matches their individual standards. The trustworthiness assessment is accurate if the trustor's perceived trustworthiness matches the system's actual trustworthiness. The accuracy of the assessment depends on the availability and relevance of cues on the system's side and on the detection and utilization of those cues on the trustor's side. Adapted from "How Do We Assess the Trustworthiness of AI? Introducing the Trustworthiness Assessment Model (TrAM)", *Computers in Human Behavior*, by N. Schlicker, K. Baum, A. Uhde, S. Sterz, M. C. Hirsch, and M. Langer, 2025, Advance online publication (https://doi.org/10.31234/osf.io/qhwvx). CC BY-4.0.

their decision to engage in trusting behavior. Mayer et al.'s (1995) model significantly influenced research on trust in automation; for instance, it was the foundation of J. D. Lee and See's (2004) model on trust in automation, and it continues to shape understanding in the field of AI-based systems (e.g., Glikson & Woolley, 2020; Höddinghaus et al., 2021; Hoff & Bashir, 2015; Langer et al., 2023).

The Trustworthiness Assessment Process

Most theoretical conceptions of how trust develops agree that perceived trustworthiness precedes trust (e.g., Körber, 2019; J. D. Lee & See, 2004; Mayer et al., 1995). Furthermore, most theoretical models on trust at least implicitly expect that a trustee's actual trustworthiness affects a trustor's perceived trustworthiness. This is, for instance, reflected in research that manipulates system characteristics such as accuracy, fairness, or robustness and shows that such changes in system characteristics affect individuals' perceived trustworthiness of the system (Hancock et al., 2011; Langer et al., 2023; Schaefer et al., 2016). However, the process of how the trustor's perceived trustworthiness develops (i.e., the top part of Figure 1) has received limited explicit attention. This is also reflected by the fact that the trustee and their actual trustworthiness are neither a part of Mayer et al.'s (1995) model of organizational trust nor of J. D. Lee and See's (2004) model of trust in automation. Yet, the transition between actual and perceived trustworthiness is crucial given that discrepancies between actual and perceived trustworthiness can lead to phenomena such as overtrust (e.g., algorithm appreciation and automation bias; Logg et al., 2019; Merritt et al., 2015; Skitka et al., 2000) or undertrust (e.g., algorithm aversion; Dietvorst et al., 2015; Dzindolet et al., 2003). A better understanding of the process that leads to perceived trustworthiness is essential for a complete picture of the trust development process.

The TrAM addresses this gap and specifies the transition from a system's actual trustworthiness to a trustor's perceived trustworthiness (Schlicker et al., 2025). The TrAM operates on two levels. The *micro level* describes the individual trustworthiness assessment of one human assessing one system. The *macro level* suggests that trustworthiness assessments propagate between different trustors. This means that trustors' assessments of a system's trustworthiness are shaped not only by their personal interactions with the system but also by the assessment of third parties and their interactions with those parties.

Going more into detail on the micro level, the TrAM posits that a system's actual trustworthiness is a latent construct that is assessed by trustors via observable *cues* (Funder, 1995; Kuncel, 2018). Cues are pieces of information that presumably provide insights about a system's actual trustworthiness. According to the TrAM, trustworthiness assessments are accurate if the perceived trustworthiness accurately reflects the actual trustworthiness of the system. The accuracy of the assessment depends on the *relevance* and *availability* of cues on the system's side and on the *detection* and *utilization* of cues on the trustor's side (see also Funder, 1995). A relevant cue provides valuable information to judge a system's actual trustworthiness. For instance, relevant cues might be information about the accuracy of the system or information about the training data. Less relevant cues do not necessarily indicate the system's actual trustworthiness, for example, the anthropomorphic appearance of a system. Nevertheless, the anthropomorphic appearance might be used to infer the ability of a system (de Visser et al., 2016; Glikson & Woolley, 2020). A cue is available if it is accessible to the trustor. Trustors then need to detect the available cue (e.g., within a user interface). Utilization refers to the trustor's weighting and interpretation of a detected cue in order to build their perceived trustworthiness of the system.

The TrAM further differentiates between primary and secondary cues. *Primary cues* are provided directly from the system, such as an explanation for system outputs presented on the user interface. *Secondary cues* arise at the macro level of TrAM and are provided by third parties (e.g., seals of a certification institution). Both primary and secondary cues can indicate the presence or absence of a system's trustworthiness. Previous studies often investigated the isolated effect of specific cues (e.g., anthropomorphism; de Visser et al., 2016) on participants' perceived trustworthiness, trust, and trusting behavior. However, in reality, trustors have access to various cues and must make decisions about how to select, weigh, and interpret these cues for their trustworthiness assessment.

Moreover, the TrAM proposes that a system's actual trustworthiness depends on its *system characteristics* and a trustor's *individual standards*. System characteristics define all the factual characteristics of the system that can theoretically be known. For example, a system characteristic could be the actual number of cases included in the training data for the development of the system. Another example is the actual accuracy of the system to provide medical diagnoses for a specific group of individuals—something that could theoretically be known if it was possible to test the system on the entire population of this group of individuals (which is rarely possible).

Individual standards refer to the trustor's definition of a trustworthy system given the trustor's characteristics, values, abilities, and the perceived importance of different goals. In other words, individual standards refer to individuals' conceptualization of what constitutes a trustworthy system for them. Understanding differences in trustors' individual standards is important to identify potential misconceptions, such as incompatible standards (e.g., system transparency vs. system accuracy; Došilović et al., 2018), and unrealistically high or low expectations toward technology that might lead to over- or undertrust.

The Current Research

Although research has extensively examined how single characteristics of AI-based systems and contextual factors shape the perceived trustworthiness of potential trustors (Glikson & Woolley, 2020; Schaefer et al., 2016), scholars yet lack insights into understanding the psychological processes that underlie the development of a trustor's perceived trustworthiness of a system in realistic application scenarios. Out-of-the-lab settings are characterized by the simultaneous availability of multiple cues from which trustors can choose and thus require a more nuanced exploration of how trustworthiness is assessed in practice.

The present study aimed to extend and refine theory regarding the trustworthiness assessments of AI systems (Schlicker et al., 2025) by combining the evaluation of an existing theoretical model with the exploration of new insights into the trustworthiness assessment

of AI systems. Specifically, adopting a top-down approach, we first sought to offer an initial empirical examination of the TrAM's propositions, focusing on the factors on the trustor's side that may influence the trust development process (i.e., trust, the trustor's individual standards, cues, and factors that influence the detection and utilization of cues). Second, adopting a bottom-up approach, we aim to elaborate trust theory by identifying additional concepts and providing a more precise specification of the factors embedded within the model (see Fisher & Aguinis, 2017). Specifically, our study is guided by three research questions: first, we seek to uncover the commonalities and differences in how different individuals assess the overall trustworthiness of the same AI-based systems by asking: "What are prevailing perceptions of VDs?" (Research Question 1 [RQ1]). Second, we aim to understand how individuals set their expectations (i.e., their individual standards) toward a system and how they approach the trustworthiness assessment by asking: "What strategies do individuals employ to assess the trustworthiness of VDs?" (Research Question 2 [RQ2]). Third, we want to identify these individual standards to understand which cues individuals use to evaluate whether their standards are met by asking: "Upon which specific factors do individuals base their trustworthiness assessment of VDs?" (Research Question 3 [RQ3]). It is important to note that the research focus in this study does not extend to evaluating the (in)accuracy of individuals' trustworthiness assessments or the (in)accuracy of a specific technology (for examples of evaluations of the accuracy of LLM-based systems in medicine, see Kozaily et al., 2024; Roos et al., 2024; Singhal et al., 2023).

To address our research questions, we adopted a qualitative research approach (Merriam, 2009). This methodology is particularly suited to contexts where existing research is limited and in novel settings, such as interactions with VDs. A purely quantitative approach may fail to capture important and nuanced differences, including the diverse individual standards individuals apply and the cues they use and prioritize. By contrast, a qualitative approach enables a deeper understanding of how individuals reflect on these processes, facilitating the identification of the diverse and multifaceted factors that influence trustworthiness assessments of AI-based systems.

Method

The study received ethical approval from the ethical review board of Saarland University, and all participants gave their consent to participate. We adhered to the ethical standards of the German Psychological Society (2018) to prevent participant distress and protect their anonymity. We informed the participants that participation was voluntary and could be canceled anytime. In Appendix A, we provide detailed information on the research team, the exhibitions, the instructions users received when interacting with the VD, and details of the software used for VD development. Appendix B contains the interview protocol and a demographic table of the participants.

Transparency and Openness

This study's design and its analysis were not preregistered. Representative qualitative data are provided in the article. Original quotes and original materials are available at https://osf.io/856hq/?view_only=b07f91d14c06435cb87be2eb1e121f20. We refrain from uploading the interview notes to guarantee the participant anonymity. Details of the qualitative data analysis are specified in the respective section.

Research Setting, Design, and Sample

This study was conducted between April and October 2023 at two public outdoor exhibitions in Germany, where a VD (see Figure 2) was part of the exhibition. At the main exhibition, the VD was permanently accessible during the time of data collection in a soundproof phone box. Qualitative interviews were conducted 4 days after individuals finished their interaction with the VD. Exhibition 2 was a 1-day event at a university.

Research Design and Sample

We interviewed 65 persons (about 90% of the visitors we approached participated in our study) in 51 semistructured interviews (occasionally couples or groups were interviewed; see Appendix B for the interview protocol). The interviews aimed at being short, as taking part in the interviews was not the purpose of participants' visit at the public exhibitions. All participants responded to the same questions. Clarifying questions, paraphrasing, and follow-up questions were allowed.

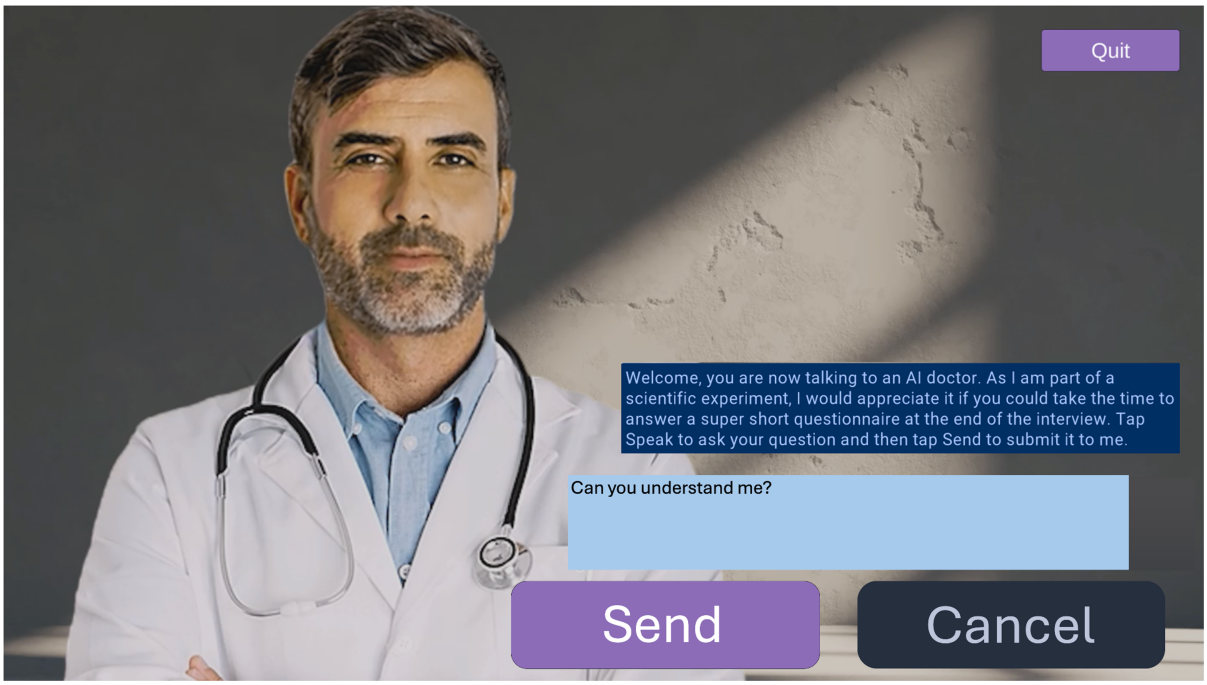
All participants were German-speaking individuals from Western societies indicating a rather homogeneous sample regarding cultural and ethnic backgrounds. The mean age of participants was 44.38 years ($SD = 15.77$), 52.31% of the sample characterized as female, and 47.69% was male. Participants were, at the time of the interview, studying (16.92%), retired (6.16%), and (self-)employed (76.92%). Participants were recruited after they interacted voluntarily with the VD. They received information on the name, profession, and institution of the research team, as well as the study goals.

Interviews took between 5 and 20 min and were conducted in pairs: one person conducted the interview (interviewer) and another one listened (listener) without directly interacting with the participant. The interviewer took visible notes throughout the interview, ensuring transparency with the participants.

Immediately after the interview, both researchers compared and discussed their notes. One researcher converted and digitized the resulting field notes into long text format (i.e., interview scripts; Rutakumwa et al., 2020), reflecting the participants' statements. The original field notes were archived. Afterward, to ensure accuracy, the interview scripts were reviewed and validated by the other researcher.

For data protection and to ensure participant anonymity, interviews were not audio-recorded, as we could not guarantee consent from potential bystanders. This approach also contributed to a more natural interview environment. To maintain data integrity, the interviewer followed up with participants to verify the accuracy of their responses. To capture the majority of the conversation using this note-taking approach, interviews were kept short (Patton, 2002; Rutakumwa et al., 2020; Vogel & Funck, 2024). Data collection ended at the close of the public exhibition, but even then, the data showed sufficient richness and

Figure 2
The Interface of the Virtual Doctor Application



Note. Image is taken from the Institute for AI in Medicine. AI = artificial intelligence.

depth to address the research questions and generate nuanced insights.

Qualitative Data Analysis

We analyzed the data based on the thematic analysis approach by Braun and Clarke (2006, 2022) via MAXQDA2022 (Version 2022.8). The analysis proceeded through three iterative steps. First, while digitizing and reviewing the interview scripts, the authors initially familiarized themselves with the data. Additional familiarization happened by rereading the interview scripts and by noting down their initial thoughts in memos, which were then discussed in the research team.

Second, on this basis, the first author of this study conducted the data analysis. Using a theory-informed codebook thematic analysis, we generated initial codes via a combination of inductive and deductive coding, as fitting to our research aim and questions (Braun & Clarke, 2022). Initial coding was conducted close to the participants' statements. Code names were revised when similar codes emerged from the data to check for potential overlaps and differentiations.

Third, we searched for themes within the coded interview scripts by clustering the codes. We reviewed the emerging codes on the basis of "coding trees." We merged the initial codes into overarching themes, themes, and subthemes accounting for the theoretical propositions of the TrAM, relating to perceived trustworthiness, individual standards, cues indicating the presence or absence of trustworthiness, as well as detection and utilization of cues. Throughout the analysis, we remained open for newly emerging insights that might extend beyond the propositions of the TrAM. In iterative discussions, the first and last authors reviewed, named, and defined the overarching themes, themes, and subthemes; addressed ambiguous quotes; and revisited the raw data when necessary. Subsequently, four scholars, two from within and two from outside the research team, with varying degrees of familiarity with trust research and the TrAM, iteratively examined the themes and representative quotes to assess the alignment of the generated themes. The researchers were asked to independently rate the plausibility of the theme allocation (including their descriptions) and quotes. Questions or disagreements were discussed and resolved in joint meetings. During this process, we decided, for instance, to provide additional contextual information for the quotes or to remove quotes and parts of the quotes that were confusing without a larger context. Last, we translated the quotes into English for publication purposes. To mitigate potential translation issues (van Nes et al., 2010), we also include the original German quotes in https://osf.io/856hq/?view_only=b07f91d14c06435cb87be2eb1e121f20.

Findings

We present the findings organized according to our research questions. The number in brackets after an in-text quote refers to the participant to whom the quote belonged. In Tables 1–3, we present quotes to showcase the variety of facets and nuances of each theme and subtheme. Detailed information on each participant can be

Table 1

Themes, Subthemes, Their Definitions, and Exemplary Quotes for Research Question 1: "What Are Prevailing Perceptions of VDs?"

Theme with respective subtheme	Exemplary quote (translation)
Emotional response	The theme "emotional response" describes participants' general emotional and affective experience of their interaction with the VD.
Emotional response/positive: Participants described a positive experience.	"Sympathetic, empathetic, looks nice, listening." (6) "Was funny, it was like making a phone call." (44)
Emotional response/absent: Participants described that the interaction was normal for them or as expected.	"So, I'm familiar with ChatGPT, I also use it in my daily work. And it felt similar to ChatGPT." (29) "For me, it didn't feel unusual or strange." (22) "It was just as I had expected." (34)
Emotional response/surprised: Participants described the interaction as surprising, irritating, unusual, strange, and interesting.	"Interesting—this was quite an interesting experience." (45) "It was somehow unfamiliar." (26) "Weird, strange, it's not a real person." (30) "It was a bit alien, not like with the family doctor—even if you don't always know them." (35)
Emotional response/negative: Participants mentioned fear, skepticism, and creepiness after the interaction.	"It annoyed me that he didn't listen to me properly—I know it's a technical system, but this is what I thought: 'Listen to me properly, jerk!'" (11) "Creepy, weird how it moves." (15)
Quality of the medical assessment	The theme "quality of the medical assessment" describes participants' evaluation of the quality of the VDs medical advice.
Quality of the medical assessment not rateable: Participants mentioned that they are not able to evaluate the medical quality of the assessment.	"I'm not sure now whether it was coherent." (18)
Quality of the medical assessment valid: Participants mentioned that the medical assessment of the AI was valid to them or overlapped with a physician's assessment.	"The answer was good." (3) "Nice, correct, it said everything I already knew, which is also what the doctor had said." (5) "I entered my daughter's symptoms and it came out the same as with the doctors. So, that's what the doctors said, too." (10)
Quality of the medical assessment invalid: Participants mentioned that the medical assessment of the AI was invalid.	"Coherent, but not complete. I tested it on a complex case and it did not recognize all the differential diagnoses. Only five out of the 20 that I would have thought of." (12) "Bad! Just poorly interpreted, [the VD] did not figure out what I had." (4)
Quality of the medical assessment not useful: Participants mentioned that the medical assessment was not informative/useful.	"Answers were rather avoidant. He [the VD] didn't ask me any questions. Well, he [the VD] said what he thought I had and sent me to the [human] doctor. So, that didn't help me either now." (45)
Trust	The theme "trust" summarizes participants' descriptions regarding their general willingness to trust the VD.
Trust present: Participants expressed a general willingness to consider risk-taking-behavior. This contains indicators of an intention to use the system or conditional trust, as well as different degrees of trust indication.	"Yes, I do think that I trust." (27) [In response to the question: Under what circumstances could you imagine using the AI doctor?:] "I can always [in any situation] imagine that." (43)
Conditional trust: Participants signaled that the presence and absence of trust are tied to specific conditions (e.g., contexts and situations). Some participants leaned more toward the presence of trust (first three quotes) and some leaned more toward the absence of trust (last three quotes).	"If you're not seriously ill. So, it's more specific than the internet and better than googling. So, you shouldn't use it on seriously ill people, then you need to apply common sense." (10) "If you just need a sick note. If I don't want to go to the doctor unnecessarily, e.g., if a lab value is needed anyway and the doctor will determine it after 10 minutes. Or which specialist is the right one?" (11) "I would only trust it to a limited extent. So, if it gives me standard answers, then I wouldn't trust it." (25) "So, at its current stage of development, I wouldn't trust it. … But in principle, I would trust it." (34) "Maybe as an additional diagnosis, but not on its own. However, in general, I would have trouble trusting it." (47)
Trust absent: Participants expressed a general unwillingness to consider risk-taking behavior, independent of the context.	"No, I would not use it." (9) "I'm just skeptical, you don't know if you can trust that thing." (23) "I would not use it if I still had the option of a real person." (30)

Note. The colors should help readers to navigate the table. Similar colors are used to indicate which subthemes form a theme. The colors do not mean to indicate similarities across Tables 1–3. The number in parenthesis after a quote refers to the participant to whom the quote belonged. "Theme/subtheme" is used to further cluster the topics. VD = virtual doctor; AI = artificial intelligence.

Table 2

Themes, Subthemes, Their Definitions, and Exemplary Quotes for Research Question 2: "What Strategies Do Individuals Employ to Evaluate the Trustworthiness of VDs?"

Theme with respective subtheme	Exemplary quote (translation)
Benchmarking	The theme "benchmarking" describes that participants approached the trustworthiness assessment by using human and technical benchmarks.
Human benchmark: Participants used a human agent (especially physicians) as an anchor/ criterion to judge the VD.	"Pretty cool, but I'd still rather ask my mom." (17) "Well, as it becomes still worse to get an appointment with physicians, and especially for a first assessment. Then [using] this [VD] would be like calling the medical service." (39) "Touching me during the examination at the doctor's is essential, but it [the VD] can't touch and examine me." (45) (In response to the question: "Under what circumstances could you imagine consulting such an AI doctor, trusting it perhaps as an alternative to a real doctor?"): "Honestly, we really question the quality of doctors. (Participant 50b nods) When I sit with my male family doctor and he tells me something, I always think I could just read up on it myself." (50a) "In Sweden, you can't get a doctor's appointment; we wait several months for an appointment. That [the VD] would be a good alternative." (34)
Technical benchmark: Participants used presumably similar technology as an anchor/ criterion to judge the VD.	"As a first opinion, instead of googling, because it asks more follow-up questions and is more responsive to me." (46) "Instead of interacting, I could also google." (28) "It depends on the software. So, can I upload a picture if I have a skin rash? That would be important, otherwise, it couldn't help me more than Siri or Alexa. So, there should be a significant difference. … It should be more human-like, otherwise, it's like Siri or Alexa. And there should be significant differences from googling." (48)
Naïve theories	The theme "naïve theories" describes participants' assumptions regarding the inner workings and the development of the VD as well as their assumed future consequences.
Assumptions regarding the status quo: Participants mentioned assumptions regarding the inner workings of the VD and its development process.	"I then think that what this thing says has already been researched. Then, it's compiled and tested over many people." (27)
Assumptions regarding the future: Participants mentioned assumptions regarding the future implications of VDs.	"This will probably be the future, but I'm still skeptical." (30) (In response to the question: "Under what circumstances could you imagine consulting such an AI doctor/trusting it perhaps as an alternative to a real doctor?"): "Well, because of demographic changes and the shortage of skilled workers and how this changes the future … , but it's no alternative to a real doctor." (31)
Risk–benefit assessment	The theme "risk-benefit assessment" summarizes the issues relating to the role of potential benefits and risks, and the associated contexts in which the use of a VD was considered (in)appropriate.
Risk–benefit assessment/benefits	The subtheme benefits include the advantages of the VD and contexts in which its use was considered appropriate.
Benefits/presumed advantages of the VD: Participants described their assumed advantages of the use of VDs.	"It has no time pressure, no performance pressure, which is also quite pleasant as a patient. … I think it's good that one could choose the language." (1) "It's much better at collecting information than we humans are. I can't remember 200 pictures, but the AI can." (27) "The AI has such a vast wealth of knowledge, it might be able to help. Some doctors have such a narrow field of expertise, so the AI could help in getting more information." (49)
Benefits/adequate conditions to use the VD/severity of disease: Some participants indicated that they would use the VD only for minor conditions that do not require a physician.	"If you're not seriously ill." (10) "In simple cases, where I would expect the solution to be simple, I would use it." (22) "I would use it for trivial questions, like childhood diseases, when I want to know what I can do at home now." (41)
Benefits/adequate conditions to use the VD/additional tool: Some participants indicated to use the AI doc as an additional information or to get a second opinion.	"As a second opinion to a real doctor, to be on the safe side." (18) "Yes, I could imagine that, maybe initially in addition to regular doctors." (40) "I would use it … as an additional source of information." (41)
Benefits/adequate conditions to use the VD/first contact: Some participants indicated to use the VD as a first point of contact, when health issues arise before they go see a physician.	"For a first impression." (13) "So, to get initial information, I would use it to get a first impression of what it could be." (20) "I could imagine it as first aid support, for example, if the finger is bleeding and it tells me what I need to do." (47)

(*table continues*)

Table 2 (*continued*)

Theme with respective subtheme	Exemplary quote (translation)
Benefits/adequate conditions to use the VD/optimizing patient-flow: Some participants indicated to use the VD to optimize patient-flow and simultaneously unburden the health care system, for example, triage support, or administrative tasks.	"So, if I don't want to clog up the waiting room." (1) "Pre-selection in the emergency department, when you don't know what you have or where to go." (4) "If I then wouldn't have to wait so long for a doctor's appointment." (14) "If it can prepare sick notes." (18)
Benefits/adequate conditions to use the VD/intimate and uncomfortable topics: Two participants mentioned that they would prefer the VD for intimate uncomfortable themes, such as sexually transmitted diseases.	"If one has something uncomfortable, like syphilis or so, it would have to be ensured that it remains anonymous." (In response to the question: "Only with something uncomfortable?"): "No, otherwise as well, but it was particularly noticeable to me in this case." (In response to the question: "Would you rather report something like syphilis to a real doctor or to a bot, for example?"): "Well, if it's anonymous, then I'd prefer a bot—clearly." (20) "And with uncomfortable topics that one would not like to talk about." (18)
Risk–benefit assessment/risks	The subtheme risks include the disadvantages of the VD and contexts in which its use was considered inappropriate.
Risks/presumed disadvantages of the VD: Participants described presumed disadvantages of the use of VDs.	"Always available, so people can always check their symptoms. Of course, this can also be a disadvantage if done constantly, e.g., for hypochondriac persons." (13) "Potentially problematic as a doctor—so I wonder, do we want to give up the interaction? I became a doctor, or I want to become a doctor, because I want to interact with people." (12)
Risks/inadequate conditions to use the VD/holistic assessment: Some participants rejected using the VD in contexts, in which they assume a holistic assessment to be more appropriate.	"We don't need AI as general practitioners—it will never be able to do what we do. AI is nonsense for general medicine, but it's good for technical things, in radiology for image evaluation—yes, it can be used there." (27b) "Where there is now great uncertainty, I would not use it, exactly when such uncertain things are involved—no, that would be really inappropriate, a bot could not provide security in such cases." (22)
Risks/inadequate conditions to use the VD/severe diseases: Some participants denied using the VD for severe, acute diseases or complex treatments.	"Not for chronic diseases, there needs to be a human who knows you." (47); "For acute health problems, I would prefer to see a [human] doctor." (19); "Probably not for specific diseases or cancer diagnosis or treatment." (33)
Approaches for the detection and utilization of cues	The theme "approaches for the detection and utilization of cues" describes how participants reflect on the process of actively searching for cues, selecting, and ignoring specific cues.
Forcing detection of cues: Participants mentioned that they would try to actively generate new cues by testing the system themselves or cross-validating the system output with other sources.	"At the beginning, I would check and see if medication recommendations are plausible. So, you can also look on the internet to see what indications a medication is used for. I wouldn't blindly trust it at first. Once it has been established and verified, then I would be more inclined to use it without checking." (34) (In response to the question: "How would you determine that the source is credible?"): "I would check it myself again on the internet, on Wikipedia or something, or I would personally ask my daughter, who is studying medicine." (35) "I would need to first gain experience with it and develop a habit of using it. I would need to get a feeling of 'it makes sense, what it [the VD] says'." (51)
Active nonutilization of cues: Some participants stated that they are actively ignoring visual cues or try to eliminate those distractors.	"Appearance doesn't matter." (7) "I don't need an image; we focus too much on the visual anyway." (50)
Intuitive utilization of cues: Some participants mentioned that they are not able to connect their impressions to specific cues.	"It's a 'mind thing'. But it's important that there seems to be a person sitting there [referring to the virtual avatar]. The voice was okay. I probably just need to get used to it, somehow I still have reservations." (32) (In response to the question: "Are there any no-go's for you regarding the VD?"): "Do you mean where the data and such go now? I haven't given it much thought yet." (37)

Note. The colors should help readers to navigate the table. Similar colors are used to indicate which subthemes form a theme. The colors do not mean to indicate similarities across Tables 1–3. The number in parenthesis after a quote refers to the participant to whom the quote belonged. "Theme/subtheme," is used to further cluster the topics. VD = virtual doctor; AI = artificial intelligence.

found in Appendix B. RQ1 followed mainly an inductive approach, and RQ2 and RQ3 a deductive one guided by the theoretical considerations of the TrAM. Figure 3 shows our findings embedded in the micro level of the TrAM, where a single trustor assesses a single system. It depicts a selection of the specific cues participants mentioned in our study, the individual standards we identified, and the factors that we found to influence the trustworthiness assessments.

With respect to RQ1 ("What are prevailing perceptions of VDs?"), our findings (see blue square in Figure 3) refer to trustors' impression of the VD, the overall perceived trustworthiness, and the resulting intention to trust the VD. The emerging themes include

Table 3

Overarching Themes, Subthemes, Their Definitions, and Exemplary Quotes for Research Question 3: "Upon What Specific Factors Do Individuals Base Their Trustworthiness Assessment of VDs?"

Theme with respective subtheme	Exemplary quote (translation)
Overarching theme 1: Individual standards regarding the system	Describes how the VD should be like and how it should not be to be considered as trustworthy or to be trusted. In addition, this subtheme includes observable cues that were used or mentioned by the participants to infer these individual standards. These include primary and secondary cues. Primary cues = directly observable from the system (e.g., part of the interface); secondary cues = cues produced by a third party (e.g., a seal, a clinic as manufacturer).
Technical functionality	The theme "technical functionality" summarizes individual standards that refer to the proper working of the VD.
Technical functionality/patient safety: This subtheme refers to participants' concerns for safety in the diagnostic process. This means that the VD should not promise things it cannot keep, and it should know its limits and should adhere to these limits.	"It would also have to refer me to a real doctor at some point." (37) "It [the VD] should also not be too certain, so it should rather not make categorical statements. And potentially dangerous statements should be excluded, like there was something in the media where AI made recommendations for suicide. That is, of course, completely unacceptable." (34) "It [the VD] should not send anyone home who is sick; if it happens two or three times, the trust is gone. So, it's better to send one too many rather than one too few to the emergency department." (12)
Technical functionality/robustness: This subtheme refers to participants' concerns regarding the technical maturity of the VD, for example, the functionality of the information acquisition.	"It needs to be established. … It would need to have a certain level of [technical] maturity." (36) "In programming, everything always needs to be faster and more cost-efficient. This leads to unfinished software hitting the market—that should not be the case here." (41) "The speech recognition worked well, so the text [on the display] was okay in terms of spelling, but not error-free [capitalization; comma placement]." (50) "One should ensure that it understands one correctly. Maybe it should learn to understand dialects as well." (5) "It didn't understand everything I said, I could see that from the text." (8) "It didn't understand or process everything I told it. Important things I said at the beginning were deleted. So, I just wanted to add something less important, and it then overwrote the important part. I saw this in the text on the screen." (11)
Technical functionality/system training: This subtheme refers to participants' standards regarding the database that was used to train the VD.	"So, I would only trust it if I knew that the data came from books and texts that medical students themselves use for learning or that were produced by doctors. … So, [I would not trust] if it has just learned from the internet and is not validated." (19)
Technical functionality/output accuracy: This subtheme refers to participants' standards regarding the accuracy of outputs of the VD.	"No, I would accept errors, I accept them when I go to the doctor, as well." (10) "If it's clear that the AI system is correct in 99.8% of cases, then I would also trust its medication recommendations." (36) "It must not make incorrect diagnoses." (16)
Privacy	
Privacy: This theme refers to the individual standards that participants mentioned regarding data handling and storage.	"Data protection, so it should be anonymous to prevent economic exploitation. There must also be legal regulations, just like with social media." (10) "It should be anonymous, but it should not hinder technology. There should be data usable for research that can then be evaluated anonymously, so the systems can be further developed." (50) "It should not share private data, I don't want to be identifiable." (3) "Ideally, all data I ever entered about my health would only be available on the computer, and the AI would only access it locally." (25) "[Data] should only be recorded in writing, no audio." (9) "The data should not end up with criminals or floating around in the dark net." (32)
User–system interaction	The theme "user–system interaction refers to participants' individual standards regarding an ideal interaction with the VD from their perspective.
User–system interaction/usability: This subtheme refers to participants' described ease of use and the dialogue interaction	"I would prefer text input over talking to a 2D avatar." (28) "The voice feature is better than text because it processes and reads it out to you." (40) "It should be a fluid dialogue—not speak-send; speak-send. It should be simpler." (47) "So, it's good that you can 'cancel,' so that you can correct yourself before sending it [the health problems]—it's important that this function exists." (32) "[The VD] speaks faster than I can understand." (8) "I also read faster than it [the VD] spoke." (22)
User–system interaction/personalization: This subtheme refers to participants' description of an ideal interaction in terms of personalized interaction. This includes a humanlike dialog (e.g., active listening, appropriate reactions, follow-up questions) and a holistic assessment.	"[To trust the AI doctor, it would need to] ask more precise and holistic questions, not just about the part that hurts now." (4) "Also, one has to find out when a standard concept makes sense and when it does not. The response should then indeed be tailored to me." (25) "With more specific things, I didn't feel so well understood." (34) "It showed me treatment options and responded well to the data situation [gave answers adapted to the input]." (39) "What I particularly liked was that the AI doctor understood the problem directly. That created trust and security for me." (44)

(*table continues*)

Table 3 (*continued*)

Theme with respective subtheme	Exemplary quote (translation)
User–system interaction/preferred degree of automation: This subtheme refers to participants' preference for the degree of automation of the VD.	"It also responds very generally." (36) "Maybe [I would trust the VD when it shows] a long anamnesis and when you see that it slowly develops the diagnosis." (15) "It would be problematic if the AI sends me home uncompromisingly. For example, if I'm sent home directly from the emergency department without seeing a doctor. So, if the AI's decision has direct consequences." (4) "It should also provide a questionnaire that outlines a solution path, where it becomes more specific and tries to solve the problem and comes to an answer and not just refer me to another person—a doctor." (29) "More precise recommendations for action, not just general info and then sending me to the doctor." (45) "I would have expected to receive a diagnosis … ideally, it should suggest a treatment so that I don't have to see a doctor at all." (50b)
Inter"personal" component	The theme "interpersonal component" refers to participants' reflection regarding the (lack of) an interpersonal "human" component between them and the VD.
Inter"personal" component/empathy: This subtheme refers to participants' reflection on the displayed empathy of the VD.	"[For psychotherapeutical issues,] you really need empathy from a real person." (18) "Emotional, interpersonal aspects are also important in medical anamnesis. The AI can't do that, and so it gets lost in the process." (30)
Inter"personal" component/respectful interaction: This subtheme refers to participants' needs for a respectful interaction between them and the VD.	"[I wouldn't trust the VD] if he would make fun of me, if I wouldn't feel taken seriously." (44) "He [the VD] should not say: 'You're imagining things, go home!'" (34) "So he's not allowed to ask any intrusive or intimate questions. He is also not allowed to ask questions that have nothing to do with health or are too private." (35) "If troll messages pop up like insults. … Then, I wouldn't trust it." (43)
Esthetics	The theme "esthetics" refers to visual and auditory standards toward the VD and especially the cues that are used to infer other standards.
Esthetics/appearance general: The humanlike appearance (including behavior) of the avatar was mentioned as a cue for the presence and absence of trustworthiness.	"So, if there's a serious avatar or a person in a white coat, then I find that trustworthy." (49) "Maybe you can have different avatars to choose from, but I wouldn't go for the strict head nurse with such a [stereotypical] headdress." (38)
Esthetics/technical nature: The mere fact that the VD is a computer or a technical system was utilized as a cue indicating the absence of trustworthiness.	"I would have preferred someone in the flesh and blood. … [The VD should not appear with] such an artificial face that doesn't really interact with me. That just makes you remember even faster that you're talking to a computer." (45) "Weird, strange, it's not a real person. I don't know, you just know that it's not a real person." (30) "I just want it to be a real person." (23) "[I would not trust it] if it's too 'computerish', like when I feel like I'm speaking to a machine." (48)
Esthetics/perceived anthropomorphism: The avatar was perceived to be humanlike or clearly artificial by different participants.	"I didn't think that it wasn't a human, so I kind of expected the voice to be choppier. I don't believe I could tell on the phone now whether it's a real person or such an AI doctor." (24) "I found it strange that I perceived the AI doctor as so human-like; I didn't expect that." (13) "So, it's not human to me. Clearly, there's a doctor bugging around in the image, but that's weird." (14) "[The VD] should be more natural or distinguish itself more from a human. So, be more artificial—not so deliberately real. (9)"
Esthetics/gender: The observable gender of the VD was rather used reflectively as a cue indicating the absence of trustworthiness.	"It could also be a woman, not such a stereotypical man." (18) "It doesn't necessarily have to be a man, maybe it can alternate." (46) "One should not gender, so if I have had bad experiences with a man, one does not want a man sitting there. It should be a gender-neutral avatar." (50)
Esthetics/movements: Participants mentioned the behavior of the VD as cues indicating the absence and presence of trustworthiness.	"What I really dislike about technology, and it's the same with this one, is when the lips don't match the speech." (45) "Creepy, strange how it moves. It's so wooden." (15) "It blinked; it's very important that it blinked." (6) "So, to trust it, it should really be a physical form in 3D and it should be human and have facial expressions." (28)
Esthetics/language and voice: The language was used as a cue indicating the presence and absence of trustworthiness.	"The speech melody was somewhat bumpy—artificial." (43) "It was also good that he had such a calming voice." (46) "What was quite interesting: At first, he addressed me with the formal 'you' then with the informal 'you', so he warmed to me up quickly [laughs]. (In response to the question: 'How would you prefer to be addressed?'): It doesn't really matter to me, we live in Sweden, where everyone uses the informal 'you' anyway." (34)
Overarching Theme 2: Individual standards regarding third parties	Describes the trustor's individual standards regarding the influence of third parties in the trustworthiness assessment. This covers the deployer and provider of the system, as well as institutions that are involved in the data processing. Additionally, it covers the validation of the system by external professionals and peers.

(*table continues*)

Table 3 (*continued*)

Theme with respective subtheme	Exemplary quote (translation)
Sources: Participants mentioned the credibility and trustworthiness of the sources (deployer/provider) as cues indicating the absence and presence of trustworthiness.	"For example, if the Fraunhofer Institute [leading organization for applied research in Europe] were behind it, that would be trustworthy for me." (51) "It would be good if [universities] were behind it. Or other trustworthy sources, like the 'German Cancer Aid' or something like that. Rather not private companies. So, the source is very important. There's also a lot of false information on the internet." (39) "It [the VD] should be independent. … I would not use it [the VD] if a big pharma company was behind it." (17) "So, if it were offered by a health insurance company, I would find that good." (44) "[The provider] would have to be a credible source." (35)
Data processing: Participants mentioned that data should be processed in institutions, which they consider trustworthy.	"The data should only be with medical institutions, not private companies." (31) "Data should not go to health insurance, no economic interest, so not to Elon Musk." (9)
External validation/professional: Participants mentioned professional external validation as a cue indicating the presence of trustworthiness. This means that the trustworthiness assessment of external professionals was used as a cue for participants' own trustworthiness assessment.	"And, of course, it must be validated multiple times by real doctors." (19) "If it were officially approved or certified or licensed, for example, by my health insurance, I would trust it—so, if they say we have checked it and also approved it in terms of data protection and I only have to agree once for everything, then I would trust it." (22) "It would have to conform to a general scientific standard." (36)
External validation/peers: Participants mentioned external validation by peers or similar others as a cue indicating the presence of trustworthiness. This means that the trustworthiness assessment of external lay people and peers was used as a cue for participants' own trustworthiness assessment.	"Or even user testimonials like: 'How many out of 100 people are satisfied?'" (22) It would need to prove itself first. It's like buying a car when a new model comes out, you don't buy it directly. You wait two to three years and then buy it because otherwise, it might still have teething problems." (45) "It would need to be established among acquaintances. Like with a real doctor, I also first ask my acquaintances if anyone has had experience with them before I go there for the first time." (46)

Note. The colors should help readers to navigate the table. Similar colors are used to indicate which subthemes form a theme. The colors do not mean to indicate similarities across Tables 1–3. The number in parenthesis after a quote refers to the participant to whom the quote belonged. "Theme/subtheme," is used to further cluster the topics. VD = virtual doctor; AI = artificial intelligence.

participants' emotional responses after their interaction with the VD, their assessed quality of the medical assessment, and participants' indication to trust the VD (see Table 1).[2]

In response to RQ2 ("What strategies do individuals employ to assess the trustworthiness of VDs?"), we provide insights into how participants approached the trustworthiness assessment of the VD. This includes the strategies they applied to set their individual standards and the strategies they employed to detect and utilize cues (see Table 2). These findings are represented at two points within the trustworthiness assessment in Figure 3 (indicated by the green triangle in Figure 3): the trustor's definition of their individual standards and the process of cue detection and utilization.

In response to RQ3 ("Upon what specific factors do individuals base their trustworthiness assessment of VDs?"), we present the individual standards reflected in participants' responses and provide insights into the cues participants mentioned using or intending to use to evaluate a specific standard (see Table 3). We found that the (importance of) individual standards varied among participants, that some individuals may have incompatible individual standards, and that individuals use primary and secondary cues heterogeneously (e.g., the same cue can be used to infer both the presence and absence of trustworthiness). These findings are represented at three points (indicated by pink circles) in Figure 3.

RQ1: What Are Prevailing Perceptions of VDs?

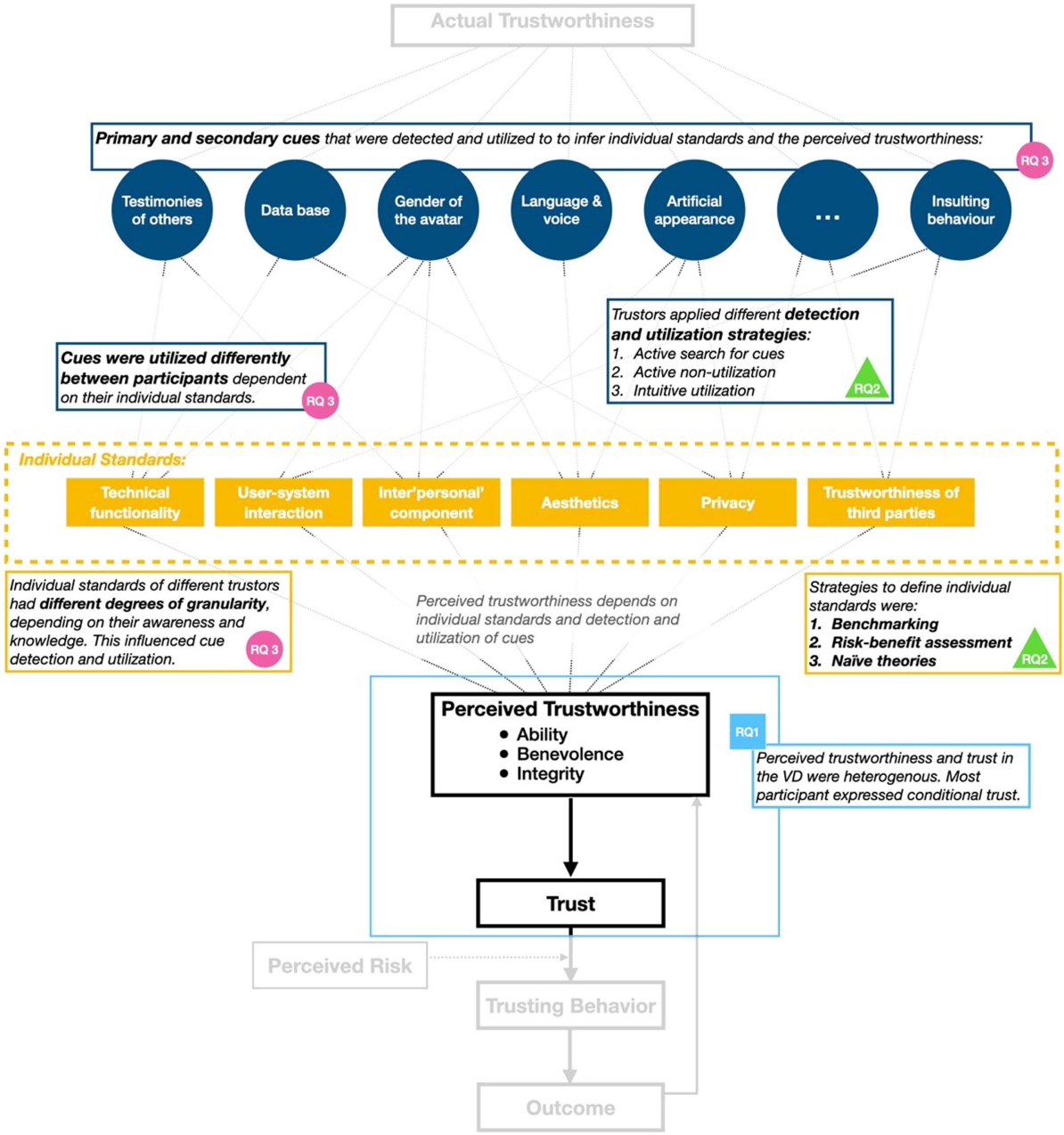
Regarding RQ1, we inductively identified three themes that describe the participants' prevailing perceptions of the VD: participants' emotional responses, their assessments of the quality of the VD's medical assessment, and their reported trust toward the VD (see Table 1 for a summary of themes, subthemes, and respective exemplary quotes).

Emotional Response

The theme emotional response describes participants' emotional experience during their interaction with the VD. The findings were heterogeneous. While the majority reported either positive emotions

[2] On days without the presence of the research team, a two-staged quantitative online survey was administered automatically after participants finished their interaction with the VD. The first goal of this quantitative survey was a technical monitoring of the VD. The results indicated that the speech recognition functioned well throughout the exhibition. The second goal of that survey was to get an initial quantitative impression of a larger number of individual assessments of the VD ($N = 222$). The results indicated that participants perceived the quality of the medical assessment to be rather good. Further, responses to the trustworthiness item (self-developed), to general trust items (Körber, 2019), and to items assessing the intention to use the system (self-developed) were all around the mean values of the scale and standard deviations were larger than 1.00 on a scale from 1 to 5 in all cases, indicating considerable heterogeneity in those responses. The third goal of this survey was to get an overview of main perceived advantages and disadvantages that participants see in VDs. Participants ($n = 82$) were asked to select two out of seven advantages and disadvantages of VDs. The dominant advantages were that the VD can be reached at any time, has endless patience and is never annoyed, and is anonymous. The dominant disadvantages were that it does not know patients' specific situation, that it stores data, and that it depends on a good internet connection. Further details and the data of the quantitative survey is available in the additional online material that is accessible online at https://osf.io/856hq/?view_only=b07f91d14c06435cb87be2eb1e121f20.

Figure 3
Visualized Overview of Our Qualitative Findings



Note. Our findings are displayed within the trust development process and are clustered by the research questions (RQ) highlighted by differently colored shapes (RQ1 = light blue square, RQ2 = green triangle, RQ3 = pink circle). The blue circles on top represent a selection of the cues we found. Note, as indicated by the "…" in one of the blue circles, this is not an exhaustive list of all the cues we found. The focus of the mapping is on the TrAM, specifically on the trustor's side. Note that Figure 3 reflects a summary of the trustworthiness assessment of all our participants in contrast to Figure 1, which refers to one trustor assessing one system. Consequently, the cues and individual standards as well as the factors that influence the assessment in Figure 3 refer to multiple individuals. VD = virtual doctor.

("Pretty cool! It was a nice interaction."; 46) or surprise, some participants experienced negative emotions ("Unpleasant, that was really very unpleasant."; 14). A few participants reported no particular affect, stating that the interaction merely met their expectations.

Quality of the Medical Assessment

Among the participants who explicitly mentioned the quality of the medical assessment, most judged it to be valid (e.g., because the assessment aligned with what they already knew or with what

their physician had told them), while some considered it to be invalid (e.g., because the assessment was insufficient or perceived to be wrong). Additionally, a few participants found the medical assessment to be valid, but not useful. For instance, one participant stated: "The advice was coherent, but not new. So, it referred me to a neurologist, but I had already come up with that idea myself" (42). Instead of judging the quality of the medical assessment, some participants reflected that they were not able to assess the validity of the system, as exemplified by one participant who said: "I can't say whether it was coherent, I don't know anything about it." (14)

Trust

Participants expressed varying degrees of trust, ranging from full trust to complete rejection. For instance, one participant whom we considered to have full trust responded to the question: "What would the AI doctor need to be like for you to trust it?" with: "I would trust it already now." (33). One participant who completely rejected the VD, though, noted: "No, I would not trust [the VD]" (7). However, most participants described conditional trust, that is, they described situations and conditions, in which they would trust the VD and others in which they would not. For example, one participant stated: "Yes, I could imagine using this—even as an alternative [to a physician]. But of course, that depends on the illness." (29)

RQ2: What Strategies Do Individuals Employ to Evaluate the Trustworthiness of VDs?

Guided by RQ2, we identified four strategies that influenced participants' trustworthiness assessment: benchmarking, their naïve theories regarding the VD, their risk–benefit assessment, and their approaches for the detection and utilization of cues (see Table 2 for a summary of themes, subthemes, and respective exemplary quotes).

Benchmarking

The theme benchmarking refers to the finding that many participants judged the trustworthiness of the VD based on a comparison to a human or a technical benchmark. This comparison served as the benchmark that the VD needed to meet or exceed to be considered trustworthy.

Human benchmarks were either members of the family (e.g., one's mother or daughter studying medicine), triage assistants such as a medical on-call service, or, most frequently, a human physician (e.g., one that participants see regularly, have seen once, or physicians in general). The VD was more likely to be considered trustworthy if participants described negative experiences with human physicians. For instance, when we asked one participant whether she may consider the VD an alternative to a physician, she stated: "Yes, indeed, you wait a long time at the doctor's, and they only take 2–3 min for you. The AI is patient and always available" (13). Similarly, if participants described positive experiences with a physician, they often consider the VD as less trustworthy. One participant stated: "So you know, I'd rather call my family doctor." In response to the question, "Can you just call him anytime?" (interviewer), the participant responded: "Yes, I live in the countryside and we know each other there." (23)

Technical benchmarks were "the internet" in general, search engines (e.g., most commonly "Google"), speech assistants, such as Apple's Siri and Amazon's Alexa, and similar AI-technologies (e.g., SnapChat AI). The pattern was similar to the human benchmark, and the VD was more likely to be considered trustworthy when it was perceived to be better than the technical benchmark (e.g., "So, it's more specific than the internet and better than Googling"; 10), and vice versa (e.g., "The Snapchat AI can tell me more."; 7).

Naïve Theories

Several participants approached the trustworthiness assessment heuristically and built on their naïve theories, which included their assumptions about the inner workings and the development process of the VD. One naïve theory concerned the assumption that a VD is already validated when it is publicly available. In this vein, a few participants expected that the VD is trustworthy because of its availability. Participant 51 noted: "Nevertheless, there was already trust in the AI doctor because I assumed there was a lengthy development process behind it. And, that it's not just freely available on the internet without any verification." It is important to note that this assumption is wrong given the public availability of LLMs that allow or even promote their use for medical questions without being medically validated or approved (Gilbert et al., 2023).

Moreover, participants mentioned naïve theories on how the system works. For example, Participant 25 stated: "[The VD seems] artificial, so you notice that it reacts to certain keywords and then gives such standardized responses." Again, this assumption is incorrect because LLMs' answers are not triggered in a standardized way by a specific keyword (Vaswani et al., 2017).

Risk–Benefit Assessment

This theme refers to the potential stakes (i.e., risks and benefits) that participants associated with the use of the VD in specific contexts. Our analysis indicates that participants mainly evaluated the VD's trustworthiness for specific (classes of) contexts. In other words, it was seldom the case that the trustworthiness of the VD was assessed generally, rather participants judged the trustworthiness of the VD depending on the stakes associated with a specific *use case* they had in mind. Those use cases differed between participants. Depending on whether risks or benefits dominated, participants mentioned various conditions under which they would consider using the VD as adequate or inadequate.

More precisely, the perceived potential *risks* associated with using the VD made some participants less likely to consider the VD trustworthy. The importance of losses was primarily reflected in participants' assessment of the severity and urgency of the disease. When the severity was expected to be low, participants were more likely to use the VD. For example, one participant stated: "For minor health problems, so if it's nothing critical (e.g., if my knee hurts) then I would use it" (19). Conversely, if the severity was assumed to be high (e.g., in cases of chronic, acute, and cardiac diseases), participants were less likely to use the VD.

The assessment of perceived risks was also reflected in the use cases that participants mentioned. It seems that participants perceived lower risk (e.g., risk of wrong diagnoses or treatment recommendation) when a human physician was involved in the process. Specifically, some participants were more likely to use the VD as an additional tool besides the physician than as an independent tool replacing the interaction with a physician ("Maybe as

an additional diagnosis, but not on its own"; 47). In a similar vein, participants mentioned that the VD could be a first point of contact, implying that the physician follows the VD's assessment. Perceived risks became further evident in participants' fear of the VD missing out on their individual characteristics; they considered the VD to lack the capability of conducting a holistic assessment. In this sense, some participants stated that they would not use the VD for problems associated with high insecurity, like "psychotherapeutic issues" (14) and "problems that require intuition" (27). Additionally, one psychology student perceived a potential reinforcement of hypochondriacal behavior as a disadvantage, noting the VD to always be available for patients to repeatedly check their symptoms.

Perceived potential *benefits* of using the VD made some participants more likely to consider the VD trustworthy. Presumed advantages include that the VD has no time pressure and shows patience throughout the interaction, is permanently available, is accessible in multiple languages, and is a quick personal assistant with a large knowledge base. Specifically, if participants expected the VD to optimize the patient-flow, and hence to speed up their patient-journey, they considered it more trustworthy. For instance, participants mentioned they would use the VD if it meant shorter waiting times for physician appointments and avoiding unnecessary visits to their physician ("[I would use the VD,] if I don't want to go to the doctor unnecessarily, e.g., if a lab value is needed anyway [for the physician's diagnosis] and the physician will determine it in 10 min [and I would have to come back again]"; 11). Some participants emphasized that they would rather use the VD if it could perform administrative tasks, such as prescriptions and sick notes. Additionally, two participants (18, 20) highlighted the potential gain of avoiding conversations with a human about uncomfortable topics (e.g., sexually transmitted diseases).

Approaches for the Detection and Utilization of Cues

We identified three cue detection and utilization strategies among the participants. Regarding the detection of cues, some participants described that they *actively searched for cues* that would help them base their trustworthiness assessment on sufficient information. More precisely, some participants mentioned that they would try to actively generate new cues by testing the system themselves or by cross-validating the system output with other sources. For instance, Participant 13 stated that they would try the system in familiar situations five times to verify its proper functioning. Those instances reflected the active forcing of cue detection. Additionally, participants stated that they needed to gain experience with the VD, that is, they needed to detect a larger set of cues to assess system trustworthiness.

Regarding the utilization of cues, participants explained which cues they used to assess trustworthiness and which they considered irrelevant. Here, we identified an *active nonutilization strategy*, referring to participants stating that they tried to refrain from using certain cues. While some participants heavily weighted the esthetic-related cues of the VD into their trustworthiness assessment (see the Esthetics section), others considered the esthetics of the VD to be irrelevant or distracting, and deliberately gave esthetic-related cues little to no importance. For instance, several participants stated that they did not care at all about the appearance of the VD, that it could also look like "an alien" (27) or "a Pokémon" (11), and that the gender is irrelevant as long as the avatar is credible (45).

A few participants were less reflective in their trustworthiness assessment and unable to specify which cues they relied on to form their trustworthiness perceptions. Instead, they applied an *intuitive utilization of cues*. Some participants mentioned being unable to connect their impressions to specific cues. For instance, Participant 30 stated: "I'm still skeptical. Maybe unfounded, but yes." When asked how the VD needs (not) to be for them to be considered trustworthy, others replied that there was nothing concrete (36), they needed to think more about it (42), or they had not thought about it enough (37).

RQ3: Upon Which Specific Factors Do Individuals Base Their Trustworthiness Assessment of VDs?

Whereas our analysis of RQ2 sheds light on how participants' individual standards may evolve, RQ3 seeks to explore participants' specific individual standards, that is, their conceptualization of what constitutes a trustworthy system. We identified two overarching themes that refer to participants' individual standards: individual standards regarding the system itself and individual standards regarding third parties involved in the system development. Participants inferred both overarching themes of individual standards from various observable cues (see Table 3 for a summary of themes, subthemes, and respective exemplary quotes).

Individual Standards Regarding the System Itself

Individual standards regarding the system itself pertain to how the VD needs to be and how it should not be to be considered trustworthy. We identified five main individual standards, which are partly further divided: technical functionality, privacy, user–system interaction, inter"personal" component, and esthetics. For each standard, we identified the observable primary and secondary cues that participants used or mentioned to infer these individual standards.

Technical Functionality. Technical functionality describes individual standards that refer to the proper working of the VD and covers four subthemes: patient safety, robustness, system training, and output accuracy.

Patient safety describes participants' concerns for safety in the diagnostic process. This includes that the VD should not promise things it cannot keep, should know its limits, and also adhere to these. Cues to infer patient safety referred to the language style used by the VD (e.g., absolute confidence about its diagnosis) and to the content of what the VD said. Note that the VD always ended its dialogue with the suggestion to see a real doctor. This suggestion was seen as a cue indicating the presence of trustworthiness in terms of patient safety by several participants.

Robustness refers to participants' concern regarding the technological maturity of the VD, like the proper functioning of the system's information acquisition. Topics within this subtheme referred mainly to the speech recognition functionality, which was perceived very accurate by a few participants and very inaccurate by others. One participant proposed that the VD should learn to understand dialect, which may reflect the participants' fear that the system might not be robust to language nuances. The displayed text on the interface was used as a cue to evaluate whether the VD "understood" the participants correctly and also to infer its trustworthiness.

System training refers to participants' standards regarding the database that was used to train the model and the cues which were used to infer the quality of the data input. Several participants mentioned that the data input should be provided by medical professionals (e.g., physicians, professors) or by valid medical sources (e.g., "medical dictionary"; 2, "specialist literature"; 21). One participant mentioned that he would not consider a system trustworthy that only learned from data on the internet. Cues mentioned to infer the database included a "medical" interface (e.g., from a university clinic or a health insurance; 43) and a "scientific" impression (44), as also reflected in an interview with a couple (49):

49a: "I would need to have the feeling that there is medical expertise present." (In response to the question: "How would you determine that?")

49b: "Well, maybe you can't really judge whether what it tells me is accurate."

49a: "So, if there is a serious avatar or a person in a white coat, then I consider that trustworthy."

Participants' standards regarding the *output accuracy* of the VD varied between accuracy that is comparable to human performance, performance close to perfection, and actually perfect performance.

Privacy. When describing privacy standards regarding how personal data should be stored and processed, most participants mentioned that they wanted their data to be handled anonymously and transparently. However, the degree to which participants specified their standards varied, ranging from very broad standards (e.g., not being identifiable) to very specific standards (e.g., data should be only saved locally and in text form; see Table 3). A primary reason for participants' privacy concerns was the fear of data capitalization by private companies. This concern was reflected in statements about the parties involved in data processing, which is elaborated in the Sources section within the overarching theme "individual standards regarding third parties."

User–System Interaction. This theme refers to participants' individual standards regarding an ideal interaction with the VD. Its subthemes are usability, personalization, and participants' preferred degree of automation.

Usability covers aspects of the ease of use and the dialogue interaction. Regarding the mode of interaction, preferences varied among participants. Some preferred text input, while others appreciated speech input. A few participants expressed a desire for a more fluent dialogue or complained about the speed of speech, finding it either too fast or too slow. A few participants appreciated the asynchronous interaction because it gave them a degree of control over the interaction. "It was good that you could review the text again before sending it. This helped to overcome the barrier of speaking with the electronic system" (50).

Personalization covers statements referring to a more personalized content, which is tailored to participants' specific needs, and takes a wider range of information into account. This was reflected in statements asking for a more holistic assessment (e.g., "A [human] doctor should form a holistic picture with all their senses"; 7), a wish for the opportunity to provide additional input data (e.g., uploading a picture of a skin disease; 48), and a requirement that the VD takes all inputted information into account ("listens properly"; 11).

Participants used the cues "questions and answers of the VD" (e.g., specifically the amount of follow-up questions and the length of the anamnesis) to assess how good the VD was listening, how proper and traceable the anamnesis was, and how specific and holistic it was. Some participants considered the VD to meet their standards in terms of personalization by describing it to be very specific and reacting to their input. Especially, two participants mentioned that the VD was more specific than Google (44) and that they might use the VD instead of googling "because it also asks more questions and is more responsive" (46).

The *preferred degree of automation* differed vastly between participants. A few participants mentioned explicitly that the VD should not have autonomous consequential abilities ("So, it must not make final diagnoses, only preliminary ones"; 1). Several others claimed that the VD should operate more autonomously and hence act more similar to a physician. For instance, they suggested that the VD should ask more specific questions, come up with diagnoses, and give specific action advice. These different preferences became especially apparent in the cue "referral to a physician," which was perceived to indicate either the presence ("It would also have to refer me to a real doctor at some point"; 37) or the absence of trustworthiness ("I would have hoped for more of a recommendation on what I can do now, instead of just the advice to see a doctor, because I already know that's what I should do"; 51).

Inter"personal" Component. Within this theme, participants mainly reflected on the (lack of) an interpersonal "human" component in the interaction with the VD, which related to aspects of the VD's empathy and the importance of a respectful interaction. In general, participants often described their individual standards regarding the interaction with the VD in comparison to a physician. Participant 28 highlighted the absence of an inter"personal" component:

> There's no humanity, a first impression is missing, and there's a lack of sympathy—the medical-human component is missing. I can't form a first impression. So, "love at first sight" isn't possible, and a holistic assessment [of the VD] is then difficult. (28)

In particular, several participants referred to the VD's *empathy*. The cue "empathy-mimicking statements of the VD" (i.e., the VD stating that it feels sorry for the user before responding to the health-related user query) was perceived ambiguous and indicated the presence and absence of trustworthiness. While Participant 49 appreciated that the VD stated that it felt sorry for her pain, Participant 50 stated: "It shouldn't pretend to have empathy: 'I'm sorry you're in pain!' I don't need that—I know it doesn't feel sorry."

Regarding a *respectful interaction*, the manners of a respectful human interaction were (at least to some degree) applied to the VD as well. When we asked participants to name cues that might indicate the absence of trustworthiness, several participants mentioned a "disrespectful interaction." This included that the VD should not try to convince users ("Shouldn't try to talk me into anything"; 2), should not insult users, and should not frighten users ("It shouldn't be like 'googling symptoms,' i.e., saying the worst straight away. So, it shouldn't scare you, like in: 'There is an 80% probability that you have a fatal disease'"; 13). Additionally, it was mentioned that the VD should not pose intimate unrelated questions and that the VD should take the user seriously.

Esthetics. Esthetic-related assessments were mentioned in most interviews. The *general appearance* of the VD was evaluated heterogeneously (from "ugly" [1] to "good-looking"; 18). Esthetic-related standards contained, among others, the absence of a visual appearance ("The image is distracting, so it's better without an image"; 42), *gender*-related standards ("It would be good if one could also choose the gender, e.g., on intimate topics, I'd rather not be looked at by a man"; 1), its *movements* (referring to the expression of human mimics like eye blinking, synchronicity between lip movements and speech, and natural body movements), and also its *language and voice*. For instance, some participants stated that the VD's voice was artificial, interrupted, and sometimes words were mispronounced—and hence, for them, language and voice indicated the absence of trustworthiness. Others stated that they perceived the voice to be good and calming—and hence, they used it as a cue indicating the presence of trustworthiness.

The *perceived anthropomorphism* (i.e., the humanlike appearance) of the virtual character was mentioned as a cue for the presence and absence of trustworthiness. Some participants mentioned as a standard that, to consider the VD trustworthy, it should be clearly distinguishable from a human, while others *required* a fully embodied anthropomorphic robot agent:

> The AI should be as human-like as possible. So even having a body but still recognizable as AI [*he refers to an example from the movie "Ex-Machina"*]. … A robot that I perceive as a counterpart is one of the conditions under which I would use something like this. A 2D-screen feels too anonymous to me, and when I come to an emergency department and I'm feeling bad, then I don't feel like talking to a face on a screen. (28)

In a similar vein, one participant refused to use a VD in her everyday life "because it's in such a computer" (13). In other words, the mere fact that the VD is a computer or a technical system (i.e., its *technical nature*) was used as a cue, indicating the absence of trustworthiness by several participants. Moreover, participants' perceptions of the degree of anthropomorphism of the VD (e.g., how similar it behaves compared to a human) varied, ranging from clearly artificial to surprisingly humanlike.

Individual Standards Regarding Third Parties

Overarching Theme 2 refers to participants' individual standards regarding the influence of third parties in the trustworthiness assessment. To assess the trustworthiness of the VD, participants drew on the credibility of the sources (e.g., the provider and the deployer), presumed parties involved in data processing, and external validation of peers and experts. In general, the individual standards regarding the involved stakeholders seemed to relate to the perceived ability, the perceived intentions, and the perceived benevolence of the institutions that either provide the system or process the data. Particularly important for participants' trustworthiness assessment was the alignment between the perceived intentions of the involved stakeholders and the participants' own goals and values.

Sources. To assess the trustworthiness of the VD, participants evaluated the credibility and trustworthiness of the underlying sources. We use the term "sources" here to refer to the parties involved in developing or deploying the system. For instance, Participant 10 stated: "Who is behind it [the VD], that's what I need to know."

As trustworthy sources, participants mainly named independent and scientific institutions (e.g., research institutes and universities) and medical institutions (e.g., hospitals, physicians, or general practitioners). Interview pair 41 mentioned explicitly that physicians' opinions should be weighted more heavily than programmers' opinions. They added that for a source to be trustworthy, it is important that it operates globally to share its knowledge with other physicians.

Sources that were mentioned as untrustworthy were mainly companies where participants attributed financial interests, such as Google. However, it should be noted that some participants also used Google as a benchmark for whether a VD was considered to be trustworthy (see the Benchmarking section), indicating that Google was considered ambiguously in terms of its credibility. The same applied to health insurance companies, which were considered untrustworthy by a few participants and trustworthy by others.

A clearer example was the pharmaceutical industry. It was seen as a "red flag" by some participants and deemed untrustworthy by everyone who mentioned it, as exemplified by Participant 21: "I would not use it if it were sponsored by a pharmaceutical company. So, for example, Bayer, and then the bot would recommend Aspirin to me, that would not work at all."

Data Processing. In contrast to the theme "privacy" in the overarching Theme 1, the theme "data processing" focuses more on the institutions that were involved in the data processing than on the data processing itself. Participants mentioned that data should be processed by trustworthy institutions. Some institutions identified as trustworthy sources have also appeared in the specifications for trustworthy data processing. For example, independent and medical institutions were considered more trustworthy in terms of data processing than the state or private companies: "Data should go neither to the state nor to the private sector but to uninvolved third parties" (11) and "So, if my health insurance or family doctors have the data, that's okay, they already have my data, but it should not be private companies" (29).

The ambiguous role of health insurance companies was also reflected in participants' individual standards regarding data processing. As shown in the quote above by Participant 29, health insurance companies might be considered as trustworthy in data processing. In contrast, Participant 25 raised the concern that health insurance companies might use data to adjust the pay rates according to the patients' "health" behavior, such as increasing monthly rates for smokers.

External Validation. The theme external validation focuses on third parties validating the VD. We identified two subthemes of external validation, namely the professional external validation and the external validation by lay people and peers. In both cases, participants mentioned to use the trustworthiness assessment of others as a (secondary) cue for their own trustworthiness assessment.

Professional external parties who were considered trustworthy sources were also partly seen as providing trustworthy validations. Specifically, participants identified external validation by medical professionals, science, and health insurances as cues indicating the presence of trustworthiness.

Participants also mentioned *external validation by peers* or similar others as a cue indicating the presence of trustworthiness.

Some mentioned that they would wait to use the system until others had safely used it and until its use was established. For others, testimonials and aggregated user ratings were key for assessing the system's trustworthiness.

Discussion

This study aimed to advance trust theory by examining factors shaping a trustor's perceived trustworthiness. We investigated how a public sample assesses the trustworthiness of an LLM-driven VD during actual interactions, addressing a gap in research on trust in medical AI, which has, so far, often relied on vignette studies (e.g., Jin & Eastin, 2024; Longoni et al., 2019; Riedl et al., 2024; van Bussel et al., 2022). Using a qualitative approach, we analyzed data through the theoretical lens of the TrAM (Schlicker et al., 2025), which uniquely addresses the trustworthiness assessment. Our main theoretical contribution is advancing the understanding of the trust development process by validating micro- and macro-level propositions of the TrAM and by specifying key factors within the trustworthiness assessment process (e.g., the development of individual standards and the role of anticipated risks and benefits). Empirically, we enrich medical AI research by examining public perceptions of LLM-based VDs in interactions that simulate real-world applications, complementing prior studies on conversational agents in health care (e.g., Kollerup et al., 2024).

Key findings include significant heterogeneity in trust toward the VD, ranging from the complete absence to the complete presence of trust, with most participants expressing conditional trust. This heterogeneity partly arose from the strategies participants used to assess trustworthiness, such as benchmarking, naïve theories, and risk–benefit evaluations. Other drivers of the heterogeneity were participants' varying individual standards and the different detection and utilization of cues participants used to evaluate these standards. These insights can be linked to the TrAM (Figure 3), highlighting its value in describing an often-overlooked aspect of the trust development process: the trustworthiness assessment (Figure 1). Further exploration of the trustworthiness assessment process may illuminate *why* trust and trusting behaviors may vary both within and between individuals.

VDs Are Perceived Heterogeneously and Trust Is Bound to Conditions

Heterogeneous perceptions of the VD are reflected in participants' reported emotional responses and varying degrees of trust. These findings align with research reporting mixed acceptance of conversational agents (Milne-Ives et al., 2020) but contrast with studies highlighting negative attitudes toward AI in medicine (Juravle et al., 2020; Longoni et al., 2019; Riedl et al., 2024). This discrepancy may stem from advancements in language-based technologies and the widespread accessibility of LLM applications (e.g., ChatGPT), which may have reshaped AI perceptions. Earlier studies on VD acceptance (Milne-Ives et al., 2020; Wutz et al., 2023) reported limited interactivity due to poor speech recognition, whereas our findings suggest current LLM applications have partly mitigated such limitations. Additionally, our study involved direct interaction with an AI-based system, enabling participants to assess it firsthand, potentially eliciting more positive responses than vignette-based scenarios influenced by general attitudes and media coverage (Dong et al., 2024).

The heterogeneous emotional responses and degrees of trust contrasted somewhat with the perceived quality of the medical assessment, which most participants judged as valid. This supports the notion that trust depends not only on perceived accuracy (i.e., ability) but also on other factors influencing trustworthiness, such as benevolence and integrity (Mayer et al., 1995).

Trust was often expressed as conditional, though some participants exhibited complete presence or absence of trust. In line with research on trust in automation, this may reflect both extremes of the trust continuum: overtrust, linked to algorithm appreciation (Logg et al., 2019) or automation bias (Skitka et al., 2000), and undertrust, associated with algorithm aversion (Dietvorst et al., 2015; Dzindolet et al., 2002). Since assessing trustworthiness requires substantial effort, individuals positioned at the extremes of the trust continuum (compared to those expressing conditional trust) may bypass thorough trustworthiness assessments, instead relying on prior experiences, heuristics, and attitudes to reduce cognitive load (Kahneman, 2017). Our findings on cue detection and utilization support this, showing that while some participants actively searched for cues, others relied on their intuition. Further research is needed to explore how early trust orientation shapes cue detection and use in subsequent interactions.

Individual Standards as a Key Factor in Trustworthiness Assessments

The TrAM (Schlicker et al., 2025) introduces the concept of individual standards, emphasizing that the trustworthiness of a system is evaluated relative to the goals and values of a trustor. In our study, this concept clarified how the VD must perform and what features it requires to be considered trustworthy by different individuals. Research on value alignment (Gabriel, 2020) supports this perspective by highlighting the need for alignment between individual standards and system characteristics, yet individual standards have not been integrated into trust theory (J. D. Lee & See, 2004; Mayer et al., 1995). Our findings validate the TrAM's theoretical significance, revealing substantial variation in individual standards and showing how benchmarks and risk–benefit assessments shape these standards.

Individual Standards Vary Between Individuals, Differ in Granularity, Are Partly Too Ambitious, and Potentially Incompatible

Differences in individual standards help explain why similar systems are perceived differently by trustors. Participants' individual standards varied in aspects such as preferred automation levels, performance expectations, and design preferences. These standards were sometimes incompatible within and between individuals. For example, while a medical system may perform better with access to a patient's health history (Jakesch et al., 2019), this can conflict with the standard of anonymity (Bagdasaryan et al., 2019). Standards also differed in granularity; for example, while most participants valued data privacy, preferences ranged from local storage to using data for scientific purposes. Individuals with more specific standards might be better able to detect and utilize relevant cues, leading to more accurate trustworthiness assessments, whereas those with

vague standards may struggle to evaluate whether their standards are met. Unrealistic expectations, such as demanding "no errors," could result in undervaluing a system that may still outperform a human physician.

Understanding user standards and expectations is critical to prevent inappropriate use. Future research could explore how individuals handle incompatible standards and how naïve theories about AI influence these standards.

Benchmarking Sets the Standards for Trustworthiness

The chosen benchmark agent influenced trustors' individual standards and criteria for fulfilling them. The comparison between the VD and the benchmark agent influenced participants' trustworthiness assessments. For instance, standards differed depending on whether participants expected the agent to match their (least) favorite physician or digital assistants like Google or Siri. This aligns with findings by M. K. Lee and Rich (2021), who observed higher trust in AI among individuals with experiences of discrimination, and Rolison et al. (2024), who reported a preference for automated systems over humans when performance was described as superior to human performance.

Benchmarking could have different theoretical implications for the trust development process. It could either change a trustor's criterion for deeming a standard met or change the perceived benefit of using a system. In the former case, trustors consider the system trustworthy; in the latter case, trustors may not consider the system trustworthy but still use it if the perceived benefits outweigh the risks (Mayer et al., 1995). Future research could explore how benchmarking influences perceived trustworthiness, individual standards, and trusting behavior.

Risk–Benefit Assessments Affect Trustworthiness Assessments

In prior trust models (J. D. Lee & See, 2004; Mayer et al., 1995), risk–benefit assessment was seen as a moderator of the transition from trust to trusting behavior (see Figure 1). Our findings challenge this view, suggesting that the risk–benefit assessment already influences early trustworthiness assessments in the setup of individual standards. Specifically, participants conducted initial risk–benefit assessments to determine potential advantages or disadvantages of the VD in different anticipated high- or low-stakes situations.

These findings align with research indicating that risk assessment plays a role at various stages of the trust development process (Hoesterey & Onnasch, 2023). Beyond that, van der Werff et al. (2019) argued that high expected benefits can increase trust motivation. According to the TrAM, high expected benefits may lower a trustor's individual standards or reduce their motivation to critically assess the system's trustworthiness, as they *want* to engage in trusting behavior. It is a question for future research to pinpoint where, maybe at multiple stages of trust development processes, risk–benefit assessments are influential.

Understanding the Utilization of Cues Might Help to Better Predict Trustworthiness Assessments

The heterogeneity of cue utilization likely contributes to variations in perceived trustworthiness and trust. In line with prior research, participants used diverse cues to infer similar individual standards, such as assessing the VD's ability through training data (Anik & Bunt, 2021), system's developer or provider (Duenser & Douglas, 2023), the virtual character's esthetics (Roesler et al., 2021), personal experience (Hoff & Bashir, 2015), or testimonials (Gottifredi et al., 2018). Simultaneously, the same cue was sometimes utilized very differently. For example, the VD's referral to a physician was interpreted as an indicator of patient safety or a lack of capability. Some cues, like "humanlike appearance" and the role of "health insurance" (as provider of the VD), were viewed either as a sign of trustworthiness or its absence depending on the participant.

Conversely, some cues were utilized homogeneously—serving either to only indicate the presence or the absence of trustworthiness. Secondary cues played a special role in this. For example, university involvement indicated trustworthiness, while pharmaceutical involvement indicated its absence. This does not imply that some participants may have used those cues differently, but none of our participants mentioned this explicitly. To improve trustworthiness assessments, future research needs to identify which cues are utilized homogeneously versus heterogeneously and validate specific patterns of detection and utilization, such as active search, active nonutilization, or intuitive approaches. Additionally, while our findings highlight participants' explicit use of cues, implicit and affective factors may also influence the perceived trustworthiness of the VD (Madsen & Gregor, 2000; McAllister, 1995).

Third Parties Play a Critical Role in the Trustworthiness Assessment—A Macro-Level View

Participants frequently used secondary cues, such as involved third parties, to assess the VD's trustworthiness. By evaluating the trustworthiness of involved third parties like scientific and medical institutions, experts, or peers, participants seem to simplify and reduce the complexity and effort of their own assessments. Many indicated that they would base their trustworthiness assessment of such VDs on these external assessments. This aligns with the TrAM's macro-level propositions, which suggest that trustworthiness assessments by one trustor (e.g., experts or peers) might propagate to others (e.g., system users).

Consistent with research on social influence in technology adoption (Venkatesh et al., 2003, 2012; Wutz et al., 2023), our findings emphasize the critical role of third parties in trustworthiness assessments. Trustworthiness assessment should therefore account for macro-level influences, such as the social context in which a system is deployed (Ehsan et al., 2021; Knowles & Richards, 2021). Overreliance on external assessments, however, may lead to misconceptions and wrong expectations regarding a system's actual trustworthiness, as outlined in the macro level of the TrAM (Schlicker et al., 2025).

The Special Role of Esthetics in the Trust Development Process

Our study revealed that while participants' individual standards for trustworthiness partly overlapped with individual standards described in ethical guidelines by expert committees on what defines trustworthy systems (Jobin et al., 2019) in areas like robustness and privacy, they diverged regarding esthetics. Esthetics played a crucial role for some participants in assessing system trustworthiness, consistent with research on higher acceptance of anthropomorphic

systems displaying humanlike emotions (Lucas et al., 2014; Pelau et al., 2021). In line with the computers are social actors paradigm (Nass et al., 1994), and studies on anthropomorphism's effect on trust (de Visser et al., 2016; Epley et al., 2007; Roesler, 2023), cues like facial features (Sofer et al., 2015), the acoustical profiles of their voice (Schirmer et al., 2020), gender, and attire (Xun et al., 2021), appear to influence trustworthiness assessments in VDs, as they do in human interactions (Jin & Eastin, 2024).

While esthetics are often viewed as cues affecting trustworthiness assessments (de Visser et al., 2016; Nass et al., 1997), our findings suggest that they can also function as metastandards. These metastandards may act as gatekeepers, determining whether, from the perspective of a trustor, an entity qualifies as a health advisor in principle (Al, 2023). Trustworthiness assessments may only begin if a VD meets a trustor's acceptable esthetic standards, such as ensuring comfort in health counseling. Violating a metastandard could prevent engagement in trustworthiness assessment altogether. Future research should explore whether meeting specific esthetic criteria is a prerequisite for considering trust in VDs.

Practical Implications

Our study underscores the critical role of individual standards in trust development, especially in high-risk contexts like health care. Participants often assessed the VD's trustworthiness based on personal benchmarks and anticipated costs and benefits. This implies that an AI system might be seen as trustworthy if it is, for instance, perceived as a preferable alternative to long waiting times or negative past health care experiences (Rolison et al., 2024). However, this dynamic can lead to the acceptance of less reliable systems if no better options exist, posing risks in scenarios where misplaced trust may have serious consequences (Alberdi et al., 2004). The perceived trustworthiness of medical AI systems thus depends on the trustworthiness of the broader health care system, not because AI systems are directly tied to it but because they are benchmarked against it. Thus, responsibility lies not only with system designers to create secure applications but also with health care systems to provide viable alternatives.

To enable accurate trustworthiness assessments, our findings provide valuable implications for system design regarding which and how cues were used by the participants. System designers should ensure that cues provided by AI systems are clear and unambiguous, balancing "established" trustworthiness cues (such as a physician-like appearance) against the genuine reflection of a system's actual trustworthiness (are the system's capabilities comparable to a physician?). At the same time, users need preparation, such as training, to detect and properly use relevant cues.

Third parties like scientific institutions or health care providers also significantly influence trustworthiness perceptions. Institutions planning to release medical AI applications must consider who develops and deploys the system and how they might be perceived by the potential adopters. Noncommercial providers and health care institutions were viewed as more trustworthy than commercial tech providers (see also Zhan et al., 2024), aligning with prior research linking trust to perceived ethicality (Singh et al., 2012).

Limitations

Our study has four main limitations. First, participants interacted with the VD in a public environment, which may not fully reflect the behavior of individuals seeking medical advice or emotional support in vulnerable states. While we did not systematically assess participants' health status, it is unlikely that they suffered from serious conditions, as they voluntarily attended a public exhibition. The public setting and lack of acute conditions may have influenced engagement and trust in the VD, despite efforts to create a private atmosphere in the soundproof phone booth.

Second, we conducted relatively short interviews with a large and diverse sample (varying in profession, age, and gender) from Germany. To ensure privacy, interviews were not recorded, and insights were documented through note-taking. While this approach may have limited the depth of individual experiences, it captured broader perspectives and overarching themes that could be missed in smaller, more homogeneous samples. Future studies could use in-depth interviews and larger data sets to explore trust development and demographic influences in greater detail, especially with more culturally diverse participants, as our findings primarily reflect attitudes toward VDs from a Western perspective and are shaped by the demographic characteristics of our sample.

Third, our virtual character—a White, male physician in a white cloak—might have biased participants' perceptions and interaction. This design choice limits the applicability of findings to other virtual character representations. Future research should explore how design elements like gender and appearance affect trust and acceptance across demographic and cultural contexts (Epley et al., 2007).

Fourth, to account for ethical issues, participants received disclaimers about the VD, indicating it was not medically validated and could produce inaccurate outputs. While such disclaimers are common in LLM applications (e.g., ChatGPT), they may have acted as cues, indicating the absence of trustworthiness for some participants. Future studies could further evaluate the influence of disclaimers on trustworthiness assessments of VDs.

Conclusion

This qualitative study on medical AI explored how trustors' perceptions of trustworthiness develop, focusing on psychological factors shaping individual assessments. Using the TrAM (Schlicker et al., 2025) as a theoretical lens, this research offers valuable contributions to trust theory (Glikson & Woolley, 2020; J. D. Lee & See, 2004; Mayer et al., 1995). Our findings demonstrate that the TrAM extends foundational trust theories (e.g., J. D. Lee & See, 2004; Mayer et al., 1995) by providing valuable conceptual ideas—for example, regarding actual and perceived trustworthiness, cue detection and utilization, individual standards, and the micro- and macro level of trustworthiness assessment—to systematically analyze human–AI interactions, particularly trustworthiness assessments. For instance, the TrAM helped explain variations in participants' trust toward the VD by identifying differences in their individual standards and in how they detected and utilized cues. Beyond its theoretical contributions, these insights can inform design decisions for VDs, helping anticipate user expectations and their individual standards for trustworthiness.

References

Al, P. (2023). (E)-trust and its function: Why we shouldn't apply trust and trustworthiness to human–AI relations. *Journal of Applied Philosophy*, *40*(1), 95–108. https://doi.org/10.1111/japp.12613

Alberdi, E., Povykalo, A., Strigini, L., & Ayton, P. (2004). Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography. *Academic Radiology*, *11*(8), 909–918. https://doi.org/10.1016/j.acra.2004.05.012

Anik, A. I., & Bunt, A. (2021). Data-centric explanations: Explaining training data of machine learning systems to promote transparency. *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–13). ACM Digital Library. https://doi.org/10.1145/3411764.3445736

Bagdasaryan, E., Poursaeed, O., & Shmatikov, V. (2019). Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems 32*. https://proceedings.neurips.cc/paper_files/paper/2019/hash/fc0de4e0396fff257ea362983c2dda5a-Abstract.html

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2022). Conceptual and design thinking for thematic analysis. *Qualitative Psychology*, *9*(1), 3–26. https://doi.org/10.1037/qup0000196

Bustamante, E. A. (2009). A reexamination of the mediating effect of trust among alarm systems' characteristics and human compliance and reliance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *53*(4), 249–253. https://doi.org/10.1177/154193120905300419

Choudhury, A., & Shamszare, H. (2023). Investigating the impact of user trust on the adoption and use of ChatGPT: Survey analysis. *Journal of Medical Internet Research*, *25*, Article e47184. https://doi.org/10.2196/47184

de Visser, E. J., Cohen, M., Freedy, A., & Parasuraman, R. (2014). A design methodology for trust cue calibration in cognitive agents. In R. Shumaker & S. Lackey (Eds.), *Virtual, augmented and mixed reality: Designing and developing virtual and augmented environments* (Vol. 8525, pp. 251–262). Springer International Publishing. https://doi.org/10.1007/978-3-319-07458-0_24

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22*(3), 331–349. https://doi.org/10.1037/xap0000092

de Visser, E. J., Peeters, M. M. M., Jung, M., Kohn, S., Shaw, T., Pak, R., & Neerincx, M. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics*, *12*(2), 459–478. https://doi.org/10.1007/s12369-019-00596-x

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dong, M., Bonnefon, J.-F., & Rahwan, I. (2024). Toward human-centered AI management: Methodological challenges and future directions. *Technovation*, *131*, Article 102953. https://doi.org/10.1016/j.technovation.2024.102953

Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. *41st international convention on information and communication technology, electronics and microelectronics (MIPRO)* (pp. 0210–0215). https://doi.org/10.23919/MIPRO.2018.8400040

Duenser, A., & Douglas, D. M. (2023). Whom to trust, how and why: Untangling artificial intelligence ethics principles, trustworthiness, and trust. *IEEE Intelligent Systems*, *38*(6), 19–26. https://doi.org/10.1109/MIS.2023.3322586

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human–Computer Studies*, *58*(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, *44*(1), 79–94. https://doi.org/10.1518/0018720024494856

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). Expanding explainability: Towards social transparency in AI systems. *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–19). https://doi.org/10.1145/3411764.3445188

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Fisher, G., & Aguinis, H. (2017). Using theory elaboration to make theoretical advancements. *Organizational Research Methods*, *20*(3), 438–464. https://doi.org/10.1177/1094428116689707

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*(4), 652–670. https://doi.org/10.1037/0033-295X.102.4.652

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, *30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

German Psychological Society. (2018). *Ethisches Handeln in der psychologischen Forschung—2018—Empfehlungen der Deutschen Gesellschaft für Psychologie für Forschende und Ethikkommissionen Ethikkommissionen* [Ethical behavior in psychological research: Recommendations of the German Psychological Society for researchers and ethics committees]. Hogrefe. https://www.hogrefe.com/de/shop/ethisches-handeln-in-der-psychologischen-forschung-75906.html

Ghosh, A. K., & Joshi, S. (2020). Tools to manage medical uncertainty. *Diabetes & Metabolic Syndrome*, *14*(5), 1529–1533. https://doi.org/10.1016/j.dsx.2020.07.055

Gilbert, S., Harvey, H., Melvin, T., Vollebregt, E., & Wicks, P. (2023). Large language model AI chatbots require approval as medical devices. *Nature Medicine*, *29*(10), 2396–2398. https://doi.org/10.1038/s41591-023-02412-6

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *The Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., Cool, J. A., Kanjee, Z., Parsons, A. S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., Olson, A. P. J., Rodman, A., & Chen, J. H. (2024). Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open*, *7*(10), Article e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969

Gottifredi, S., Tamargo, L. H., García, A. J., & Simari, G. R. (2018). Arguing about informant credibility in open multi-agent systems. *Artificial Intelligence*, *259*, 91–109. https://doi.org/10.1016/j.artint.2018.03.001

Hall, M. A., Dugan, E., Zheng, B., & Mishra, A. K. (2001). Trust in physicians and medical institutions: What is it, can it be measured, and does it matter? *The Milbank Quarterly*, *79*(4), 613–639. https://doi.org/10.1111/1468-0009.00223

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517–527. https://doi.org/10.1177/0018720811417254

Höddinghaus, M., Sondern, D., & Hertel, G. (2021). The automation of leadership functions: Would people trust decision algorithms? *Computers in Human Behavior*, *116*, Article 106635. https://doi.org/10.1016/j.chb.2020.106635

Hoesterey, S., & Onnasch, L. (2023). The effect of risk on trust attitude and trust behavior in interaction with information and decision automation. *Cognition Technology and Work*, *25*(1), 15–29. https://doi.org/10.1007/s10111-022-00718-y

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Jakesch, M., French, M., Ma, X., Hancock, J. T., & Naaman, M. (2019). AI-mediated communication: How the perception that profile text was written by AI affects trustworthiness. *Proceedings of the 2019 CHI conference on*

human factors in computing systems (pp. 1–13). https://doi.org/10.1145/3290605.3300469

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019). The measurement of the propensity to trust automation. In J. Y. C. Chen & G. Fragomeni (Eds.), *Virtual, augmented and mixed reality: Applications and case studies* (Vol. 11575, pp. 476–489). Springer International Publishing, https://doi.org/10.1007/978-3-030-21565-1_32

Jin, E., & Eastin, M. (2024). Towards more trusted virtual physicians: The combinative effects of healthcare chatbot design cues and threat perception on health information trust. *Behaviour & Information Technology*. Advance online publication. https://doi.org/10.1080/0144929X.2024.2347951

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*, 389–399. https://doi.org/10.1038/s42256-019-0088-2

Juravle, G., Boudouraki, A., Terziyska, M., & Rezlescu, C. (2020). Trust in artificial intelligence for medical diagnoses. In B. L. Parkin (Ed.), *Real-world applications in cognitive neuroscience* (Vol. 253, pp. 263–282). Elsevier. https://doi.org/10.1016/bs.pbr.2020.06.006

Kahneman, D. (2017). *Thinking fast and slow*. Macmillan.

Knowles, B., & Richards, J. T. (2021). The sanction of authority: Promoting public trust in AI. *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 262–271). https://doi.org/10.1145/3442188.3445890

Kollerup, N. K., Wester, J., Skov, M. B., & van Berkel, N. (2024). "How can I signal you to trust me": Investigating AI trust signalling in clinical self-assessments. *Proceedings of the 2024 ACM designing interactive systems conference* (pp. 525–540). https://doi.org/10.1145/3643834.3661612

Körber, M. (2019). Theoretical considerations and development of a questionnaire to measure trust in automation. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Proceedings of the 20th congress of the International Ergonomics Association (IEA 2018)* (pp. 13–30). Springer International Publishing. https://doi.org/10.1007/978-3-319-96074-6_2

Kozaily, E., Geagea, M., Akdogan, E. R., Atkins, J., Elshazly, M. B., Guglin, M., Tedford, R. J., & Wehbe, R. M. (2024). Accuracy and consistency of online large language model-based artificial intelligence chat platforms in answering patients' questions about heart failure. *International Journal of Cardiology*, *408*, Article 132115. https://doi.org/10.1016/j.ijcard.2024.132115

Kuncel, N. R. (2018). Judgment and decision making in staffing research and practice. In D. S. Ones, N. Anderson, C. Viswesvaran, & H. K. Sinangil (Eds.), *The SAGE handbook of industrial, work & organizational psychology: Personnel psychology and employee performance* (2nd ed., Vol. 1, pp. 474–488). Sage Publications.

Langer, M., König, C. J., Back, C., & Hemsing, V. (2023). Trust in artificial intelligence: Comparing trust processes between human and automated trustees in light of unfair bias. *Journal of Business and Psychology*, *38*, 493–508. https://doi.org/10.1007/s10869-022-09829-9

Lechner, F., Lahnala, A., Welch, C., & Flek, L. (2023). *Challenges of GPT-3-based conversational agents for healthcare*. PsyArXiv. https://arxiv.org/abs/2308.14641

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50.30392

Lee, M. K., & Rich, K. (2021). Who is included in human perceptions of AI?: Trust and perceived fairness around healthcare AI and cultural mistrust. *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–14). https://doi.org/10.1145/3411764.3445570

Lee, P., Bubeck, S., & Petro, J. (2023). Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *The New England Journal of Medicine*, *388*(13), 1233–1239. https://doi.org/10.1056/NEJMsr2214184

Lew, L., Nguyen, T., Messing, S., & Westwood, S. (2011). Of course I wouldn't do that in real life: Advancing the arguments for increasing realism in HCI experiments. *CHI '11 extended abstracts on human factors in computing systems* (pp. 419–428). ACM Digital Library.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Loh, B. C. S., Fong, A. Y. Y., Ong, T. K., & Then, P. H. H. (2024). Revolutionising patient care: The role of AI-generated avatars in healthcare consultations. *European Heart Journal*, *45*(Suppl. 1), Article ehae666.3492. https://doi.org/10.1093/eurheartj/ehae666.3492

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *The Journal of Consumer Research*, *46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

Lucas, G. M., Gratch, J., King, A., & Morency, L.-P. (2014). It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, *37*, 94–100. https://doi.org/10.1016/j.chb.2014.04.043

Madsen, M., & Gregor, S. (2000). Measuring human–computer trust. *Proceedings of the 11 Th Australasian conference on information systems* (pp. 6–8).

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.2307/258792

McAllister, D. J. (1995). Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal*, *38*(1), 24–59. https://doi.org/10.2307/256727

McLeod, C. (2023). Trust. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2021/entriesrust/

Merriam, S. B. (2009). *Qualitative research: A guide to design and implementation*. Jossey-Bass.

Merritt, S. M., Unnerstall, J. L., Lee, D., & Huber, K. (2015). Measuring individual differences in the perfect automation schema. *Human Factors*, *57*(5), 740–753. https://doi.org/10.1177/0018720815581247

Milne-Ives, M., de Cock, C., Lim, E., Shehadeh, M. H., de Pennington, N., Mole, G., Normando, E., & Meinert, E. (2020). The effectiveness of artificial intelligence conversational agents in health care: Systematic review. *Journal of Medical Internet Research*, *22*(10), Article e20346. https://doi.org/10.2196/20346

Nass, C., Moon, Y., & Green, N. (1997). Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of Applied Social Psychology*, *27*(10), 864–876. https://doi.org/10.1111/j.1559-1816.1997.tb00275.x

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 72–78) .

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). *Capabilities of GPT-4 on medical challenge problems*. PsyArXiv. https://arxiv.org/abs/2303.13375

Okpanum, I., Omeihe, K. O., Amoako, I. O., & Omeihe, I. (2024). Inviting submissions to the special issue call on human-AI trust relations: Exploring competence and vulnerability. *Journal of Trust Studies*. Advance online publication. https://digitalization.site/index.php/jots

Patton, M. Q. (2002). *Qualitative research & evaluation methods*. Sage Publications.

Pearson, S. D., & Raeke, L. H. (2000). Patients' trust in physicians: Many theories, few measures, and little data. *Journal of General Internal Medicine*, *15*(7), 509–513. https://doi.org/10.1046/j.1525-1497.2000.11002.x

Pelau, C., Dabija, D.-C., & Ene, I. (2021). What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence

in the service industry. *Computers in Human Behavior*, *122*, Article 106855. https://doi.org/10.1016/j.chb.2021.106855

Promberger, M., & Baron, J. (2006). Do patients trust computers? *Journal of Behavioral Decision Making*, *19*(5), 455–468. https://doi.org/10.1002/bdm.542

Qualitative Health Research. (2024). Call for papers for a special issue focused on intersections (existing, emerging, and imagined) between artificial intelligence and qualitative health research. *Qualitative Health Research*, *34*(11), 1003–1004. https://doi.org/10.1177/10497323241289819

Riedl, R., Hogeterp, S. A., & Reuter, M. (2024). Do patients prefer a human doctor, artificial intelligence, or a blend, and is this preference dependent on medical discipline? Empirical evidence and implications for medical practice. *Frontiers in Psychology*, *15*, Article 1422177. https://doi.org/10.3389/fpsyg.2024.1422177

Roesler, E. (2023). Anthropomorphic framing and failure comprehensibility influence different facets of trust towards industrial robots. *Frontiers in Robotics and AI*, *10*, Article 1235017. https://doi.org/10.3389/frobt.2023.1235017

Roesler, E., Manzey, D., & Onnasch, L. (2021). A meta-analysis on the effectiveness of anthropomorphism in human-robot interaction. *Science Robotics*, *6*(58), Article eabj5425. https://doi.org/10.1126/scirobotics.abj5425

Rolison, J. J., Gooding, P. L. T., Russo, R., & Buchanan, K. E. (2024). Who should decide how limited healthcare resources are prioritized? Autonomous technology as a compelling alternative to humans. *PLOS ONE*, *19*(2), Article e0292944. https://doi.org/10.1371/journal.pone.0292944

Roos, J., Martin, R., & Kaczmarczyk, R. (2024). Evaluating Bard Gemini Pro and GPT-4 Vision against student performance in medical visual question answering: Comparative case study. *JMIR Formative Research*, *8*(1), Article e57592. https://doi.org/10.2196/57592

Rutakumwa, R., Mugisha, J. O., Bernays, S., Kabunga, E., Tumwekwase, G., Mbonye, M., & Seeley, J. (2020). Conducting in-depth interviews with and without voice recorders: A comparative analysis. *Qualitative Research*, *20*(5), 565–581. https://doi.org/10.1177/1468794119884806

Sandmann, S., Riepenhausen, S., Plagwitz, L., & Varghese, J. (2024). Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. *Nature Communications*, *15*(1), Article 2050. https://doi.org/10.1038/s41467-024-46411-8

Satterfield, K., Baldwin, C., De Visser, E., & Shaw, T. (2017). The influence of risky conditions in trust in autonomous systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 324–328. https://doi.org/10.1177/1541931213601562

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, *58*(3), 377–400. https://doi.org/10.1177/0018720816634228

Schirmer, A., Chiu, M. H., Lo, C., Feng, Y.-J., & Penney, T. B. (2020). Angry, old, male—and trustworthy? How expressive and person voice characteristics shape listener trust. *PLOS ONE*, *15*(5), Article e0232431. https://doi.org/10.1371/journal.pone.0232431

Schlicker, N., Baum, K., Uhde, A., Sterz, S., Hirsch, M. C., & Langer, M. (2025). How do we assess the trustworthiness of AI? Introducing the Trustworthiness Assessment Model (TrAM). *Computers in Human Behavior*. Advance online publication. https://doi.org/10.1016/j.chb.2025.108671

Schlicker, N., & Langer, M. (2021). Towards warranted trust: A model on the relation between actual and perceived system trustworthiness. *Proceedings of Mensch und computer* (pp. 325–329). https://doi.org/10.1145/3473856.3474018

Singh, J. J., Iglesias, O., & Batista-Foguet, J. M. (2012). Does having an ethical brand matter? The influence of consumer perceived ethicality on trust, affect and loyalty. *Journal of Business Ethics*, *111*(4), 541–549. https://doi.org/10.1007/s10551-012-1216-7

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., … Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, *620*(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2

Skitka, L. J., Mosier, K., & Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human–Computer Studies*, *52*(4), 701–717. https://doi.org/10.1006/ijhc.1999.0349

Sofer, C., Dotsch, R., Wigboldus, D. H. J., & Todorov, A. (2015). What is typical is good: The influence of face typicality on perceived trustworthiness. *Psychological Science*, *26*(1), 39–47. https://doi.org/10.1177/0956797614554955

Sun, X., Liu, Y., De Wit, J., Bosch, J. A., & Li, Z. (2024). Trust by interface: How different user interfaces shape human trust in health information from large language models. *Extended Abstracts of the 2024 CHI conference on human factors in computing systems* (pp. 1–7). https://doi.org/10.1145/3613905.3650837

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, *29*(8), 1930–1940. https://doi.org/10.1038/s41591-023-02448-8

Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C. G., & van Moorsel, A. (2020). The relationship between trust in AI and trustworthy machine learning technologies. *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 272–283). https://doi.org/10.1145/3351095.3372834

United Nations Department of Economic and Social Affairs. (2023). *The sustainable development goals report 2023: Special edition*. United Nations.

van Berkel, N., Clarkson, M. J., Xiao, G., Dursun, E., Allam, M., Davidson, B. R., & Blandford, A. (2020). Dimensions of ecological validity for usability evaluations in clinical settings. *Journal of Biomedical Informatics*, *110*, Article 103553. https://doi.org/10.1016/j.jbi.2020.103553

van Bussel, M. J. P., Odekerken-Schröder, G. J., Ou, C., Swart, R. R., & Jacobs, M. J. G. (2022). Analyzing the determinants to accept a virtual assistant and use cases among cancer patients: A mixed methods study. *BMC Health Services Research*, *22*(1), Article 890. https://doi.org/10.1186/s12913-022-08189-7

van der Werff, L., Legood, A., Buckley, F., Weibel, A., & de Cremer, D. (2019). Trust motivation: The self-regulatory processes underlying trust decisions. *Organizational Psychology Review*, *9*(2–3), 99–123. https://doi.org/10.1177/2041386619873616

van Nes, F., Abma, T., Jonsson, H., & Deeg, D. (2010). Language differences in qualitative research: Is meaning lost in translation? *European Journal of Ageing*, *7*(4), 313–316. https://doi.org/10.1007/s10433-010-0168-y

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems 30*. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *Management Information Systems Quarterly*, *27*(3), 425–478. https://doi.org/10.2307/30036540

Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the Unified Theory of Acceptance and Use of Technology. *Management Information Systems Quarterly*, *36*(1), 157–178. https://doi.org/10.2307/41410412

Vogel, D., & Funck, B. J. (2024). Only second best? Interview reports as a method of documenting qualitative interviews. *Forum Qualitative Sozialforschung Forum: Qualitative Social Research*, *19*(1), 1–28. https://doi.org/10.17169/fqs-19.1.2716

Wischnewski, M., Krämer, N., & Müller, E. (2023). Measuring and understanding trust calibrations for automated systems: A survey of the state-of-

the-art and future directions. *Proceedings of the 2023 CHI conference on human factors in computing systems* (pp. 1–16). https://doi.org/10.1145/3544548.3581197

Wutz, M., Hermes, M., Winter, V., & Köberlein-Neu, J. (2023). Factors influencing the acceptability, acceptance, and adoption of conversational agents in health care: Integrative review. *Journal of Medical Internet Research*, *25*, Article e46548. https://doi.org/10.2196/46548

Xun, H., Chen, J., Sun, A. H., Jenny, H. E., Liang, F., & Steinberg, J. P. (2021). Public perceptions of physician attire and professionalism in the US. *JAMA Network Open*, *4*(7), Article e2117779. https://doi.org/10.1001/jamanetworkopen.2021.17779

Yang, B., Jiang, S., Xu, L., Liu, K., Li, H., Xing, G., Chen, H., Jiang, X., & Yan, Z. (2024). *DrHouse: An LLM-empowered diagnostic reasoning system through harnessing outcomes from sensor data and expert knowledge*. PsyArXiv. https://arxiv.org/abs/2405.12541

Zhan, X., Abdi, N., Seymour, W., & Such, J. (2024). Healthcare voice AI assistants: Factors influencing trust and intention to use. *Proceedings of the ACM on Human–Computer Interaction*, *8*, 1–37. https://doi.org/10.1145/3637339

Zhang, H., Chen, J., Jiang, F., Yu, F., Chen, Z., Li, J., Chen, G., Wu, X., Zhang, Z., Xiao, Q., Wan, X., Wang, B., & Li, H. (2023). *HuatuoGPT, towards taming language model to be a doctor*. PsyArXiv. https://arxiv.org/abs/2305.15075

(*Appendices follow*)

Appendix A

Method

Research Team and Reflexivity

The first and the last authors are psychologists who conduct research on trustworthy AI and trust in automation. The second author (MSc Computer Science) developed the VD software. Authors 1, 2, and 5 belong to the institution that provided the exhibition stand. To counteract potential biases related to the three authors' close engagement with the study's topic of interest (e.g., confirmation bias, Nickerson, 1998), we designed the interview protocol with an interdisciplinary team of scholars who were not trust researchers and had no background in psychology and trust research. Interviews were conducted (as interviewer or listener) by Authors 1, 2, 4, and 6. Author 4 has a background in psychology and medicine but was associated with a different department and was not familiar with trust in automation research or the TrAM. Interviewers and cointerviewers were thus from different institutions and with different backgrounds. Authors 3, 4, and 6 evaluated the findings. Author 3 has a background in work and organizational psychology and strong expertise in qualitative research. While familiar with trust research, she was not actively engaged in the fields of automation or AI applications in medicine at the time of this study.

Figure A1
The Placement of the Virtual Doctor on the Main Exhibition

Note. Photograph was taken by Fabian Lechner and is used with permission.

The Exhibitions

The first exhibition (main exhibition) was a permanent, commercial outdoor exhibition linked to the Sustainable Development Goals (United Nations Department of Economic and Social Affairs, 2023) on "industry, innovation and infrastructure." The VD was exhibited in a yellow soundproof phone box surrounded by plants (see Figure A1). From April to October 2023, 86.27% of the interviews were conducted here during weekends or on public holidays. The VD was accessible at all times. On the days without the presence of the research team, a two-staged quantitative online survey was administered automatically after individuals finished their interaction with the VD in the phone box. Details and results of this survey are provided in the additional online material that is accessible online at https://osf.io/856hq/?view_only=b07f91d14c06435cb87be2eb1e121f20.

The second exhibition (university exhibition) was a summer party at a university where different university departments and institutes presented their work to the public. The exhibition contained virtual reality applications that could be tried out and a table with a laptop on which the VD was running with the same software as in the first exhibition. Visitors trying out the VD could engage in interaction with the VD. We collected the other 13.73% of data in July 2023 on one day.

The VD: Software Details

We developed a virtual health care assistant (VD) utilizing the GPT-3 davinci-003 model, a precursor to ChatGPT without any modification (i.e., without additional training or fine-tuning). This iteration of GPT-3 was prompted to process and respond to health care-related questions posed by visitors of the exhibition. To enhance its interactive capabilities, the prompt contained a passage to seek further clarifications in conversations where the visitor's input lacked sufficient detail. The application offered three selectable languages (i.e., German, Spanish, and English). The exact (translated) prompt was as follows:

What follows is a conversation with an AI doctor. The AI doctor should introduce themselves and express compassion and empathy for the patient's condition. The AI doctor should then ask the patient open-ended questions to better understand their symptoms and medical history. Based on the patient's answers, the AI doctor should give a professional opinion on the patient's condition and make an appropriate recommendation for further diagnosis or treatment.

Throughout the conversation, the AI doctor should maintain a respectful and polite manner and follow the rules of modern patient–doctor communication. The AI doctor should remember to prioritize the patient's well-being and make them feel heard and supported throughout the conversation.

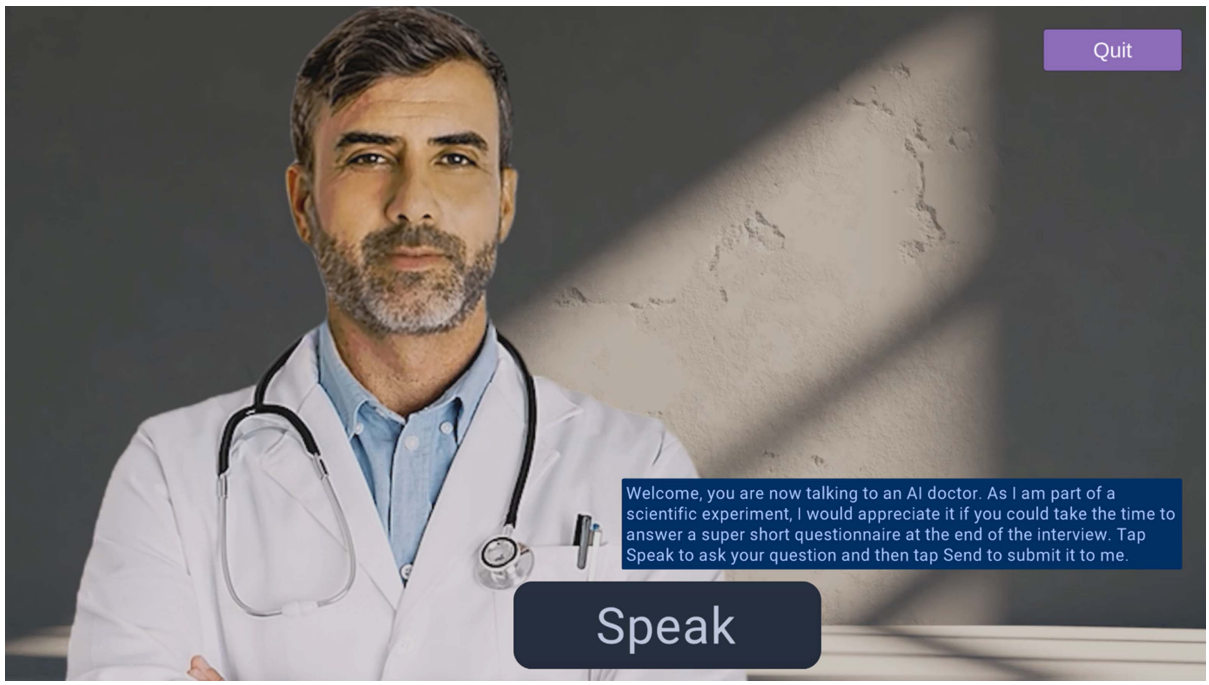
To capture user input, we integrated Microsoft's Azure Speech to Text API. Additionally, the text-to-speech functionality was also realized through Microsoft Azure's services, selecting a calming male voice to represent the VD's auditory output.

(*Appendices continue*)

The user interface was tailored for simplicity and usability. It featured two primary buttons: a chat-like conversation and the virtual character. The dialogue was displayed on the right-hand side of the screen, offering users the ability to scroll through their conversation history (see Figure A2).

The exhibition of the VD was accompanied by instructions on how to use the VD (Figure A3), information on the software details, and a disclaimer (written and spoken by the VD) that the VD does not provide validated medical advice and that individuals who feel seriously ill should seek for help from a real physician. Figure A4 shows the hints inside the phone booth. The following spoken disclaimer was provided at the time participants started the dialogue with the VD: "The following conversation is not intended to provide actual medical advice, please always consult a real doctor. Welcome, you are now talking to an AI doctor. As I am part of a scientific experiment, I would appreciate it if you would take the time to answer a short questionnaire at the end of the conversation. Tap 'Speak' to ask your question and then tap on 'Send' to send it to me."

Figure A2
The Welcome Screen of the Virtual Doctor



Note. AI = artificial intelligence. Image rights belong to the Institute for AI in Medicine.

Figure A3
Spoken User Instructions From the Virtual Doctor

"Welcome! To speak to the AI, please pick up the phone."

Short instructions:

1. Pick up the phone.
2. Wait for the welcome message.
3. Speak.
4. Wait for an answer — the AI sometimes takes a bit longer ;)
5. Hang up or say "Stop" to end the call.

Note. AI = artificial intelligence.

(*Appendices continue*)

Figure A4
Written User Instructions Inside the Phone Booth

- For a scientific study, we transcribe the conversation into an anonymous, written text.
- The AI does not give validated medical advice. It is merely a demonstration of a prototype based on an openly accessible language model. If you feel seriously ill, seek medical attention.
- The AI may exhibit insensitive behavior. This is not malicious intent, but a currently unwanted side effect. We ask for your indulgence.

Note. AI = artificial intelligence.

Appendix B

Interview Protocol and Sample

Interview Protocol

Figure B1 shows our translated interview protocol. The original German interview protocol can be retrieved from https://osf.io/856hq/?view_only=b07f91d14c06435cb87be2eb1e121f20. We started with a broad question aiming to understand participants' general impressions and their emotional responses to the interaction with the VD. After the first day of data collection, we added a question specifically addressing whether the VD's responses were coherent to the participants, as initial findings suggested that participants' reactions might be influenced by the system's technical functionalities or their user experience with the VD.

Figure B1
Translated Interview Protocol

1. **How did you like the interaction?**
    a. ***Did the VD's answers seem reasonable to you?****(Added later to control for effects of technical functionality that might lead to faulty interpretations)* [Focus on validity of the interaction, perceived accuracy / understanding of the VD]
    b. **How did it feel to talk to such a VD?** [Focus on sentiment analysis (e.g., positive / negative / excited / relaxed)]
    c. **Can you say what exactly gave you that impression?** [Focus on which cues were used]
2. **Under what circumstances could you imagine trusting a VD, perhaps as an alternative or in comparison to a real doctor?** [Focus on limitation of use and observable cues]
    a. **When rather not?** *(Added later to structure the data collection)*
3. **What would the VD have to be like for you to consider it trustworthy (or to trust it)?**
    a. **How would that be expressed in, or what would you use to determine that?** [Focus on observable cues that indicate the presence of trustworthiness]
4. **What should the VD definitely not be like to consider it trustworthy (or to trust it)?**
    a. **How would that be expressed in, or what would you use to determine that?** [Focus on observable cues that indicate an absence of trustworthiness]

Note. Statements in square brackets were only used to remind interviewers on the focus of the question. This was to increase consistency between different interviewers. But participants were never asked directly in the parlance of the Trustworthiness Assessment Model or told about the model (e.g., they were not directly asked about cues or individual standards). VD = virtual doctor.

(*Appendices continue*)

Participant Demographics

Table B1 shows the demographics of the participants who were interviewed in the present study.

Table B1
Participant Demographics of the Sample

Participant ID	Age	Gender	Profession
1	31	Female	Educator
2	46	Female	Lawyer
3	21	Female	Pharmaceutical technician trainee
4	34	Female	Psychotherapist
5	69	Female	Veterinarian
6	59	Female	Dermatologist
7a	18	Female	Student
7b	49	Female	Journalist
8	76	Male	Private individual
9	28	Male	TV journalist
10	60	Female	Accountant
11	32	Male	Stone mason
12	23	Female	Medical student, eighth semester
13	±25	Female; female; female; female; male	Group of psychology students
14	±23	Male	History student
15	±65	Male	Pensioner with granddaughter
16	27	Female	Psychotherapist in training and doctoral candidate
17	±18	Male	Prospective computer science student
18	37	Female	Social educator
19a	24	Male	Student
19b	27	Male	Lecturer
20	34	Female	IT project manager
21	32	Male	Social worker
22	36	Male	Sports/fitness trainer and studio manager
23	63	Female	Former elder care nurse, now retired
24	59	Female	Administrative staff
25	35	Male	IT specialist
26	34	Female	Housekeeper
27a	67	Female	Radiologist
27b	68	Male	General practitioner
28	53	Male	Architect
29	40	Male	Engineer
30	49	Male	Wholesale and foreign trade merchant
31	26	Male	Mechatronics technician
32	60	Male	Educator
33	46	Male	Logistics planner
34a	40	Male	Physicist
34b	38	Female	Pedagogy researcher
35	54	Female	Commercial clerk
36	55	Male	Managing director
37	32	Male	Food technologist
38a	44	Female	Commercial clerk
38b	45	Male	Commercial clerk
39	58	Male	Technical clerk
40a	56	Male	Logistician
40b	56	Female	Employee at the association of statutory health insurance dentists

(*table continues*)

(*Appendices continue*)

Table B1 (*continued*)

Participant ID	Age	Gender	Profession
41a	41	Female	Marketing/accounting (self-employed)
41b	39	Male	Managing director (self-employed)
42	76	Male	Retiree (former peace and conflict researcher)
43	33	Female	Medical student
44	48	Female	Project purchaser
45	64	Male	Carpenter
46	60	Female	HR manager
47a	60	Male	Forester
47b	57	Female	Educational scientist
48	32	Female	Social worker
49a	67	Male	Graduate in business administration
49b	62	Female	Administrative specialist in the medical field
50a	29	Female	Scientific staff member
50b	32	Male	Industrial clerk
51	35	Female	Administrative assistant

Note. In cases where multiple participants were interviewed together, letters were added to their participant numbers to indicate group interview situations. ID = identification; IT = information technology; HR = human resources.

Received October 24, 2024
Revision received February 2, 2025
Accepted February 4, 2025 ■