Trustworthy Embodied Conversational Agents for Healthcare

A Design Exploration of Embodied Conversational Agents for the periconception period at Erasmus MC

Lupetti, Maria Luce; Hagens, Emma; Van Der Maden, Willem; Steegers-Theunissen, Régine; Rousian, Melek

DOI
10.1145/3571884.3597128

Publication date
2023

Document Version
Final published version

Published in
CUI '23: Proceedings of the 5th International Conference on Conversational User Interfaces

Trustworthy Embodied Conversational Agents for Healthcare

A Design Exploration of Embodied Conversational Agents for the periconception period at Erasmus MC

Maria Luce Lupetti*
m.l.lupetti@tudelft.nl
Delft University of Technology
Delft, The Netherlands

Emma Hagens*
e.c.w.hagens@tudelft.nl
Delft University of Technology
Delft, The Netherlands

Willem van der Maden
W.L.A.vanderMaden@tudelft.nl
Delft University of Technology
Delft, The Netherlands

Régine Steegers
r.steegers@erasmusmc.nl
Department of Obstetrics and Gynaecology Erasmus MC, University Medical Center
Rotterdam, The Netherlands

Melek Rousian
m.rousian@erasmusmc.nl
Department of Obstetrics and Gynaecology Erasmus University Medical Center
Rotterdam, The Netherlands

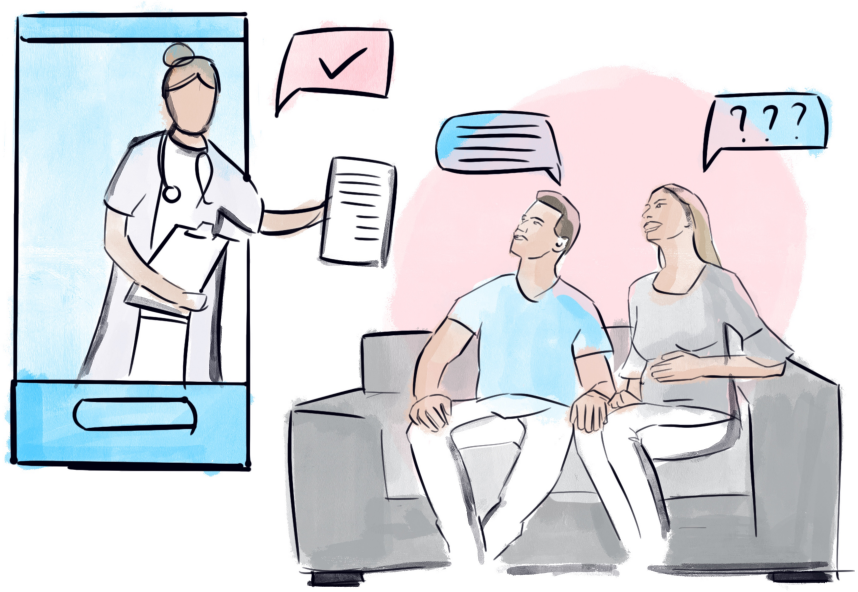

Figure 1: Patient health care journey across the periconception and pregnancy period using embodied conversational agents

ABSTRACT

This paper explores the potential implications of embodied conversational agents (ECAs) in healthcare, focusing on the impact of appearance and conversation style on trustworthiness. We conducted a Research through Design investigation of ECAs for supporting women during the periconception period and in pregnancy. The paper presents the results of a Wizard of Oz study in which two alternative prototypes, a chatbot, and an ECA, were tested in a tertiary hospital by 25 participants. Reflecting on the results we suggest that limited patients' trust in ECAs maybe be beneficial for achieving trustworthy use of these agents in the healthcare context.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in interaction design**; • **Applied computing** → **Health care information systems**.

KEYWORDS

embodied conversational agents, conversational style, trustworthiness, healthcare, research through design

*Both authors contributed equally to this research.

CUI '23, July 19–21, 2023, Eindhoven, Netherlands

ACM ISBN 979-8-4007-0014-9/23/07.
https://doi.org/10.1145/3571884.3597128

ACM Reference Format:
Maria Luce Lupetti, Emma Hagens, Willem van der Maden, Régine Steegers, and Melek Rousian. 2023. Trustworthy Embodied Conversational Agents for Healthcare: A Design Exploration of Embodied Conversational Agents for the periconception period at Erasmus MC. In *ACM conference on Conversational User Interfaces (CUI '23), July 19–21, 2023, Eindhoven, Netherlands.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3571884.3597128

1 INTRODUCTION

Current developments in artificial intelligence (AI), especially advancements in natural language processing and voice recognition, have led to a constant increase in the use of conversational user interfaces (CUI) [30], such as chatbots and virtual assistants—also referred to as conversational agents (CAs). Applications of CAs can be found in a wide variety of industries, including but not limited to customer service, entertainment, education, and healthcare, and can take the form of embodied (e.g., smart assistants) and non-embodied (e.g., chatbots) solutions. Among these sectors, healthcare has shown a particular interest in these technologies as they can be used to screen and monitor health conditions [33, 38]; support the consultation and for advice [33, 38, 62]; discussing clinical results [62]; performing triage [38]; supporting mental health [60]; providing healthy lifestyle advice [18]; and more. These applications are, potentially, of crucial importance for the healthcare sector which is struggling with an increased administrative workload and increasing staff shortages [16]. In addition, physicians are required to perform more preventive lifestyle care, to work more according to evidence-based guidelines, and to provide personalized, person-centred, care [63]. This leads to an increasing demand for adequately trained staff, who will be asked to 'do more in less time' [63], increasing the workload and negatively affecting the quality of care [40, 61, 63].

To counteract these issues, eHealth services and information delivered or enhanced through the Internet and related technologies [63]– are increasingly being adopted [56, 63] with the aim to help to lower the workload of care professionals and contribute improving health care quality [63]. CAs, especially if used to take over screening, monitoring, and advice tasks, represent a particularly promising family of eHealth technologies. These can harness some of the advantages of surveys and interviews (largely used in preventive care and prolonged care programs), such as scalability and limited costs, while also leveraging the naturalness of dialogue-based interactions [15]. Even more so, embodied conversational agents (ECAs) can further improve the acceptability and satisfaction of eHealth solutions by leveraging verbal, facial, and gestural expressions to achieve human-like procedures, such as interviews [37]. As such, a growing body of research is now invested in understanding what features, best design practices and evaluation metrics would ensure successful applications of CAs in healthcare [8, 24, 30, 38]. Nevertheless, while proven effective and possibly usable and reliable, these systems are still rarely implemented for actual use in clinical care settings, a problem that goes hand in hand with the current lack of research on the ethical impact that CAs can have in clinical care. As a matter of fact, while aspects of usability and reliability are largely addressed in CA literature, only a small number of studies explicitly investigate the ethical and trustworthiness of CAs in healthcare [3, 30], and almost none focus on the implications of embodiment [56]. Thereafter, this work investigates the potential implications of introducing ECAs in the healthcare domain in terms of trustworthiness, especially focusing on the impact of embodiment and conversational style on people's perceptions. We contextualize our work within the specific case of healthcare support for women during the periconception period (14 weeks prior to and 10 weeks after conception) [54] and in pregnancy. We developed two alternative CAs prototypes, one chatbot, and one ECA, and tested each with two alternative conditions (referencing VS non-referencing conversations) to investigate the effects of embodiment and conversational style on patients' perceived trustworthiness of ECAs. The prototypes were tested in a tertiary hospital with 25 participants who were invited during their first consultation for periconception lifestyle care at the outpatient clinic Healthy Pregnancy.

2 RELATED WORK

2.1 Conversational agents in healthcare

One of the most promising sectors for CAs deployment is healthcare. Through a variety of applications, such as screening and monitoring [25, 33, 38]; consult and advice [33, 38, 62]; discussion of clinical results [62]; triage [38]; mental health support [60]; healthy lifestyle advice [18], CAs can help to lower the workload of care professionals and contribute improving health care quality. In particular, existing research has shown that CAs are particularly suited to take over tasks where data needs to be collected systematically, usually through repeated interviews or surveys. CAs, in fact, can harness key advantages of surveys and interviews, such as scalability and contained costs, while also leveraging the naturalness of dialogue-based interactions [15]. Especially, embodied conversational agents (ECAs) can improve the acceptability and satisfaction of eHealth solutions by leveraging verbal, facial, and gestural expressions when performing human-like procedures, such as interviews [37]. Wang et al. [65], for instance, developed an ECA as a counselor in charge of documenting family health histories, which proved to be feasible, highly effective, and acceptable to participants. Interaction with ECA, and CA in general, may in fact be preferable to filling online forms [25, 37], as it can improve the understandability of the contents, especially for patients with low health literacy [25], speed up the patients' response time, and elicit open-ended responses [35]. Furthermore, as CUIs can be perceived to be free from personal biases, patients may experience less anxiety when discussing private health, especially when disclosing risky health behaviors [34].

Along with ensuring accurate [8] and successful dialogue exchanges [30], effective use of CAs in healthcare requires careful design of the agent's personality and conversation style. Preference for a certain CA personality may depend on the patient's personality and context [64], yet certain CA personality traits have been shown to have a direct effect on the user attitude and interaction. For instance, Li et al. [31] observed that, in the context of high-stakes job interviews, people are more willing to listen and confide in the artificial interviewer when this is designed as a serious and assertive agent. Within CA literature in the healthcare domain, there is a variety of personality traits that recur, such as coach-like, healthcare professional-like, informal, and more [8]. These personality traits are used to guide the development of conversation styles

that fit with the sensitive setting. In general, a formal conversation style has been observed to be perceived more positively when the conversation is about managing sensitive health information and facilitating the elicitation of more high-quality patient utterances when discussing a patient's lifestyle behaviour [15]. And more specifically, conversation style choices like calling patients by their name, starting the conversation with a social chat, using appropriate humour, providing appropriate feedback at dedicated times, and reminding information discussed in past interactions [4] have been shown to be beneficial for successful CA deployment.

Regarding the appearance of the ECAs and their impact on interactions with patients, there is still a relatively limited body of literature and no general agreement on what the best way is to represent such agents, in terms of gender, age, and rendering styles [14, 56]. Attractiveness, however, has been observed to significantly increase the agent's persuasiveness and its capacity of changing people's opinions and behaviours [22, 23, 43]. Both personality and appearance are essential aspects of the agent design as they can increase the naturalness of the interaction, resulting in more persuasive conversations [42]. Yet, persuasiveness can be a double-edged sword, especially if the naturalness of interaction results in deceiving the patient to feel and act differently than they might intend, or to trust the agent when they should not [30]. ECA designers need not only to develop appropriate appearances and conversation styles but also and foremost to define conditions and criteria for the trustworthy deployment of these agents [66].

2.2 Trustworthiness of (Embodied) Conversational Agents

As (E)CAs are increasingly being deployed to collect and manage sensitive data and to perform in delicate situations, concerns regarding their trustworthiness are rising. The first major concern raised by both patients and healthcare providers is related to the use of data. These agents have access to an increasing variety of personally identifiable information and intimate details of patients [48], who may wonder how such data is managed, and whether this is stored to be sold to marketing organizations for generating extra profit [13]. In this regard, developers and providers of (E)CAs need to ensure dignity and respect for patients [34] by designing to enable data ownership, security and privacy [11, 29, 51], in compliance with applicable regulations, such as the General Data Protection Regulation (GDPR) [48].

Another major–yet less addressed– concern is safety [30]. In case of serious health concerns, a lack of accuracy in the conversation and inconsistency of the agent's answers may have serious consequences. For instance, commercially available CAs have been shown to be inconsistent in recognizing serious health signals that may require immediate action, such as concerns about suicide, domestic violence, and rape [39]. (E)CAs should then be designed with the capability to monitor for risks automatically and then take appropriate action [34], but careful monitoring of their operation is also needed [30]. Furthermore, issues of data management and safety are intertwined with concerns regarding design biases which may exacerbate existing healthcare disparities, e.g., generating inaccurate predictions for subgroups of patients [11, 34].

While addressing some challenges emerging from the introduction of these agents in healthcare settings, such as data management and legal issues, may be beyond the scope of many design and evaluation studies [11], designers and developers of (E)CAs have the responsibility and opportunity to shape the agent's personality and appearance in a way that consciously addresses those potential risks. As research into the attractiveness of (E)CAs shows, an embodied —even more so attractive— agent is more likely to be trusted and, thus, can be more persuasive [23, 43].

Yet, a persuasive agent may deceive a patient into feeling and acting differently than they might intend [42], or to trust the agent when they should not [30] (see the case of inconsistent responses to life-threatening situations [39]). As we also learn from the related field of human-robot interaction (HRI), persuasiveness is a double edge sword: on the one hand, it allows us to achieve effective and smooth interactions, but on the other hand, may lead to people overtrust the agent [1]. Salem and colleagues [49], for instance, found that a robot showing cognitive and physical fallacies (e.g., recalling wrong user preferences, or moving erratically) is perceived as less reliable and trustworthy, yet people's willingness to comply with its instructions is not affected, 'even in the case of unusual requests'. Relatedly, Robinette and colleagues [45] observed that people tend to rely on robot guidance even when it shows poor performance and the stakes are high, e.g., in an emergency evacuation scenario. Overtrust towards artificial agents was also observed in the case of clinical decision support systems. Coiera and colleagues [12] argue that although most of these systems are accurate 80–90% of the time, occasional incorrect advice does occur but users would still rely on the inaccurate advice over their correct decision. For instance, Koppel and colleagues [27] found that most clinicians in one hospital followed computer advice to administrate drugs to patients even if the recommendations were significantly different from the dosages they used to prescribe before the system was implemented. These works exemplify a tendency of people to perceive AI-powered systems as authoritative even when lacking adequate evidence about the actual capabilities of a given system [21]. As Kapania and colleagues [21] discuss, people tend to hold AI decisions as reliable and consider AI agents as infallible and fairer than humans, up to the point of blaming themselves or other people in case of problems.

Thereafter, while the naturalness of human-like conversation and appearance may be a lever to facilitate interaction, these also need to be pondered case by case. Even more so, when developing (E)CA agents, designers need to be conscious of the difference between the patient's trust and the agent's trustworthiness (the first does not necessarily correspond to the second [26]) and account for *automation bias*— the human tendency to over-accept computer output as a heuristic replacement of vigilant information-seeking and processing [19] (the phenomenon leading to the issue of overtrust discussed above)— that has shown to be even more prominent when it comes to AI systems [21]. As lack of transparency regarding the agents' capabilities may hinder patients' ability to make informed decisions regarding their behavior (e.g., about information disclosure [47]) and influence clinicians' decision-making toward undesirable results [27], it is then of foremost importance to make obvious for people that they are talking to a machine and not led to believe that they are speaking to a human [5].

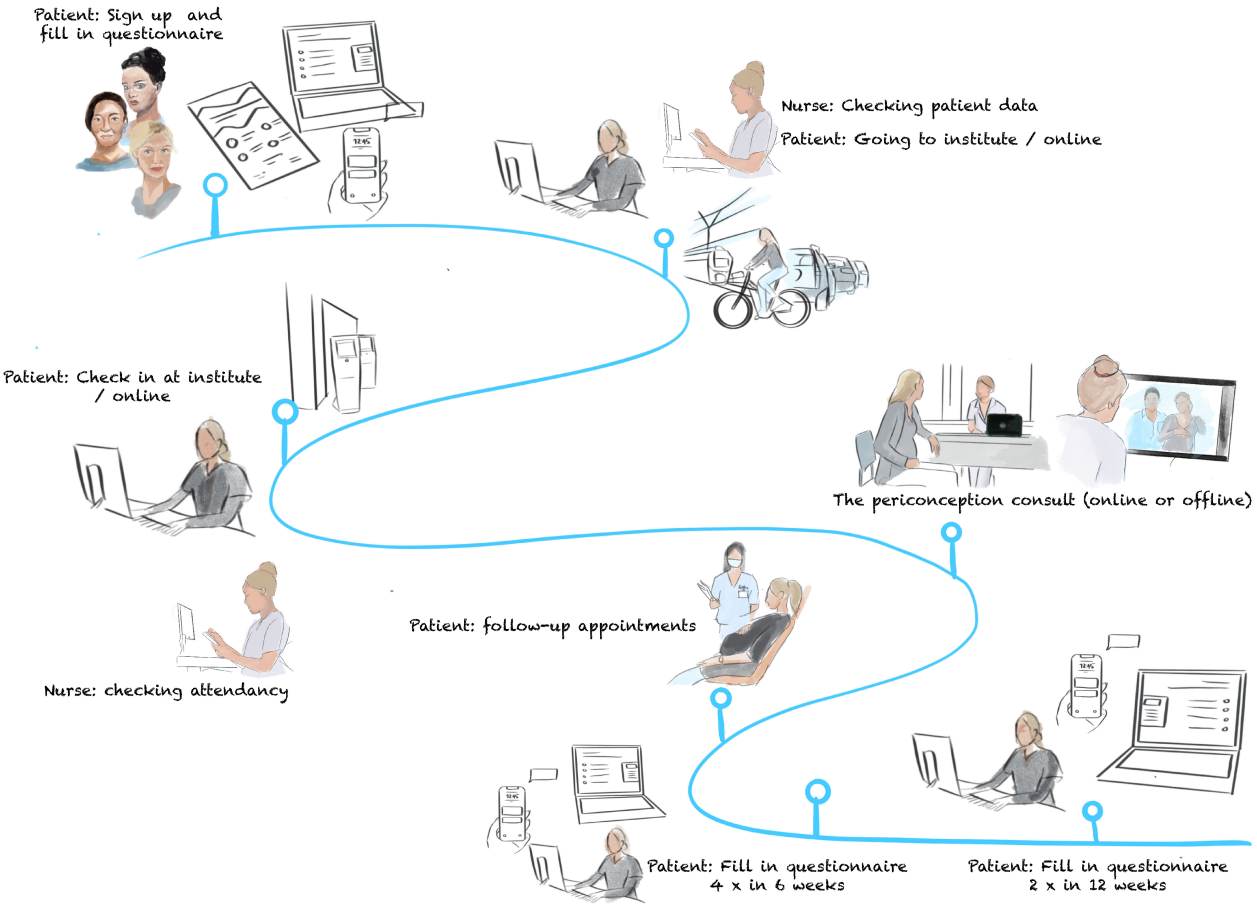

Figure 2: Journeys of nurse and patient of the care path of periconception program Outpatient Clinic Healthy Pregnancy

3 DESIGNING A TRUSTWORTHY ECA FOR PERICONCEPTION AND PREGNANCY CARE

Building on the monition from many authors to be cautious about the power of CAs persuasiveness [10, 23, 30, 43], we conducted a design investigation where we manipulated the appearance and conversation style to understand how these would impact the perceived trustworthiness of an ECA.

The team, composed of three researchers from the design field, and two from the clinical research field, designed and tested an ECA as part of a healthcare program at Erasmus University Medical Center.

The concept stems from the need of the healthcare sector to mitigate the work pressure on the healthcare personnel, such as nurses and physicians, who are increasingly asked to 'do more in less time' [63], because of staff shortages [16] and a shift towards preventive and personalized lifestyle care [63]. Specifically, our ECA is intended as an additional resource for nurses and gynecologists of the outpatient clinic Healthy Pregnancy of the Department of Obstetrics and Gynecology (OCHP) of the Erasmus University Medical Center, a facility that provides women with cutting-edge lifestyle care paths, for pregnancy and preconception [46], which is defined as the time window of 14 weeks before to 10 weeks after conception, therefore covering the vulnerable processes of gametogenesis, embryogenesis and the initiation of placentation. Our ECA is specifically designed to support the periconception part of the program (see overview in Figure 2), and to partially replace nurses: the agent only takes over repetitive and routine tasks, such as existing lifestyle questionnaires and check-in appointments. The path, in fact, includes a series of consultations and a number of systematic steps to be followed by a nurse specialist using a dedicated online platform *'Removed for Review'* [20] accessible to both healthcare specialists and patients. The aim of this path is to give patients tailor-made advice on how to promote or go through pregnancy as healthily as possible and collect clinical data.

We applied a Research through Design approach [52] to produce and preliminarily assess (with non-representative participants) a series of prototypes that allowed us to explore the implications of specific appearances and conversation styles. Given the interest in exploring the various dimensions affecting trustworthiness, we developed prototypes varying in terms of the nature of the embodiment, level of anthropomorphism, and conversational style. In

Table 1: Factors of perceived trustworthiness by Mayer et al. [36] contextualized for the healthcare context

Factors of perceived trustworthiness	Original description	description adapted to CAs for healthcare
Ability	Is the group of skills, competencies, and characteristics that enable a party to have influence within some specific domain	The ECA is capable of accurately articulating questions and interpreting answers from patients [8], providing appropriate and reliable recommendations [39], and monitoring for risks automatically and taking appropriate action if needed [34]
Benevolence	Is the extent to which a trustee is believed to want to do good to the trustor	The ECA improves healthcare programs by providing additional values for the patients [37] and ensuring their dignity [34]
Integrity	Is the trustor's perception that the trustee adheres to a set of principles that the trustor finds acceptable	The ECA collects data and provides recommendations based on expert personnel knowledge and procedures, while ensuring appropriate treatment of patients' data in compliance with data protection regulations [11, 29, 48, 51]

the first exploration, we prototyped a hypothetical routing consult conversation making use of Google Nest, adopting a referencing conversational style. Based on this preliminary exploration we gathered preliminary feedback on the choice of using a female voice and on the use of the referencing conversational style. The female voice (the voice of the researcher developing the prototype) was perceived as pleasant and appropriate for the context. The referencing conversational style was considered credible yet potentially annoying if it was not perfectly calibrated. This preliminary testing also confronted us with the fact that using commercial voice assistants would raise concerns about the potential risks in terms of privacy and data commodification. Based on this, we steered our design investigation towards the design of a dedicated agent, that would rather be placed within the existing digital environment of the hospital. To explore this direction, we developed three alternative prototypes representing the agent with different levels of visual abstraction. One consisted of a completely abstract representation where a pattern of colors would react through movement to the voice. This was perceived as too abstract and made participants feel like 'outsiders' and had little control over the behaviour of the agent. The second consisted of a photo-realistic representation of the agent as a human, generated through deep-fake techniques. The resulting images created a ghostly effect that surfaced uncanny feelings. The last prototype also consisted of a human-like agent but represented in a cartoon style, with limited sets of animations. As this was perceived as friendly, effective, and not uncanny, we used it as the ground for the final prototype.

Based on these preliminary learnings we developed one ECA and one chatbot (to be used as a comparison) to investigate whether the features we manipulated have a direct effect on the patient's perceived trustworthiness of the ECA. The two were tested in a clinical hospital with 25 participants (18 women, 7 accompanied by a male partner) taking part in the intake procedure of a periconception care program.

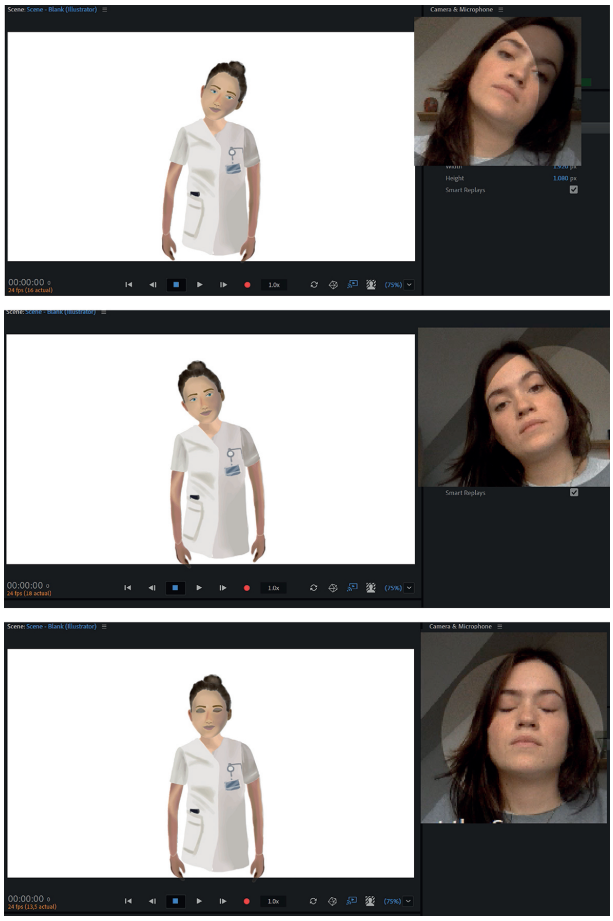

Figure 3: Head and torso movements of Robin the ECA

3.1 Conversational agent design

In order to structure our design exploration of trustworthy ECAs for healthcare, we refer to the model of trust by Mayer and colleagues [36], which proposes three factors affecting perceived trustworthiness: *ability, benevolence, and integrity*. As related works warn us about the potential counterproductivity and dangers of patients' overtrust towards CAs in healthcare [30, 39], which can be unintentionally caused when designing to encourage user trust, we also reflect on these aspects through the theoretical lens of *calibrated trust*, which we define as 'the state when a person's perceived trustworthiness of an agent matches that agent's actual trustworthiness' (adapted from de Visser et al [17]). In table 1 we provide a brief

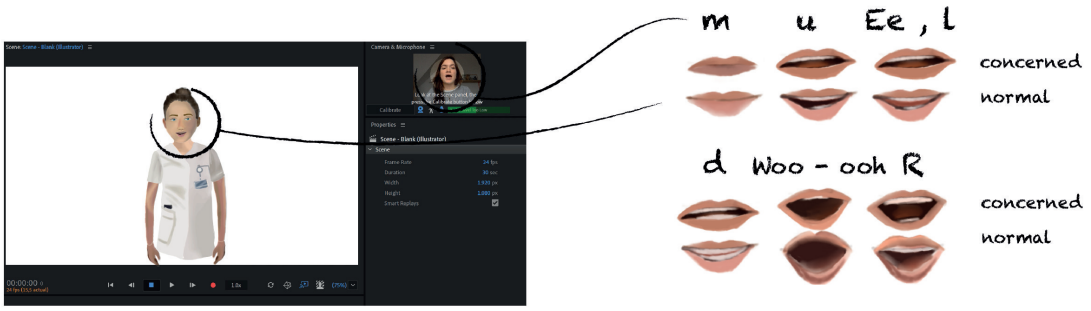

Figure 4: Phonemes of Robin the ECA

description of the three factors and a corresponding description of how we contextualize each of them within our investigation, based on previous (E)CA literature.

We used the adapted descriptions of factors affecting perceived trustworthiness (Table 1) as guidelines for our design explorations. We addressed the aspect of *benevolence*, and the need for communicating an improved experience, by focusing our intervention on interactions that are already highly repetitive, such as filling forms, which often happen with a nurse but do not actually imply (meaningful) social interaction. In previous works, in fact, we learned that ECAs are particularly suited to take over tasks where data needs to be collected systematically, such as repeated interviews or surveys [15]. Regarding *integrity*, instead, we assumed that our ECA would be perceived as respectful of patients' privacy and comply with regulations because of the high reputation of the institution we worked with. As a matter of fact, the collection of clinical data and subsequent use for research is common practice at Erasmus University Medical Center, and patients taking part in experimentations are requested to give consent for participating in the study, where they also are informed about how data is managed according to regulations. Thereafter, the main part of our design intervention is about conveying *ability*.

We altered the appearance to communicate that the ECA belongs to Erasmus University Medical Center staff and the conversation style to make clear that the suggested recommendations are based on expert knowledge, whether it is from literature or the healthcare personnel in charge of the patient care path. In this, we also embedded aspects of *integrity* by developing a procedure and related recommendations based on the expertise of the hospital personnel. Even more so, instead of developing an agent as a replacement for a nurse or obstetrician, we designed the ECA as a complementary role that would explicitly refer to the expertise of the hospital personnel when giving advice. As ensuring accurate dialogue exchanges (*ability*), monitoring risks (*ability*), and proper data treatment (*integrity*) are challenges that require interventions at a technical level, we excluded them from our exploration and focused on features that could be altered and controlled through design interventions, mentioned above.

3.2 Robin: Embodied Conversational Agent for Periconception and Pregnancy Care

We developed our ECA by iteratively testing multiple prototyping platforms and testing different approaches to the development of the appearance, mostly varying from abstract to highly human-like representations. Our final prototype is *Robin*: a cartoon-like female ECA, that we live-animated using Adobe Character Animator. We opted for a *cartoon-like* representation style to leverage the human-like capacity to provide familiar and relatable interaction, which contributes to building trust [43], but we also limited the level of realness to avoid potential feelings of eeriness and uncanniness [58]. We engaged in discussions about whether to attribute gender to our ECA, as CA's gendering is widely problematized, e.g., because people tend to attribute negative stereotypes to female CAs significantly more than to male CAs [6] and may exacerbate gender discrepancies [55]. Yet, we relate to previous research that has shown how the implications of gendering CAs depend on the context [43]. Thus, we conceptually and visually positioned *Robin* as part of the periconception care staff, which is almost exclusively composed of women (in the case of our case study).

The ECA is distinctively characterized by a white lab coat, that looks like the one worn by the clinical personnel of the Erasmus University Medical Center. Previous research, in fact, has found that an ECA presented as a healthcare professional, rather than with a casual appearance, is perceived as more reliable and authoritative, and people are more likely to follow its advice [57]. Finally, we designed the conversational style of *Robin* to match the idea of this being a complementary role (an important aspect we identified for *integrity*) and to enable patients to understand the limits of the agent (as a way to mitigate *automation bias*. We build on the work of Kowatsch and colleagues [28] who demonstrated how a CA designed as a personal assistant of a healthcare professional can be highly acceptable and lead towards a strong working alliance between the agent and the patients. As in their work, we designed for *Robin* a *referencing* conversational style. The ECA introduced itself as the personal assistant of a healthcare professional by mentioning his/her name, and, several times during the intervention, it provided feedback and recommendations explicitly referring to the healthcare professional.

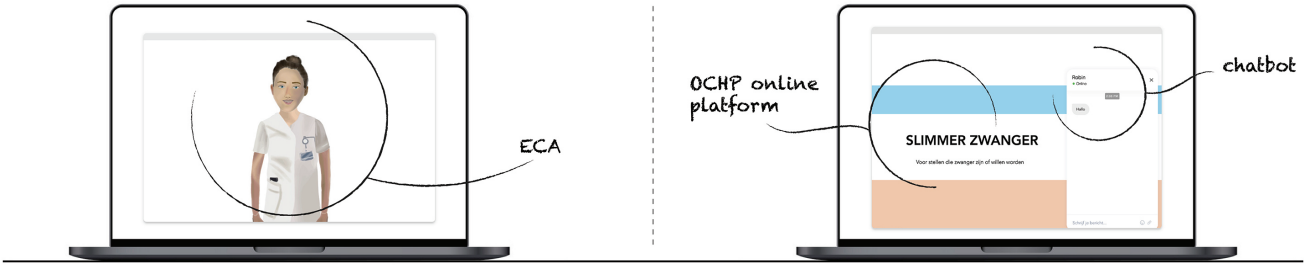

Figure 5: Conversational agents used in the study, presented on a screen. On the left, the Embodied Conversational Agent. On the right, the Chatbot

4 STUDY DESIGN

We employed our ECA within real consults attended by women participating in the periconception care program, some of whom were accompanied by their partners. Our intervention was intended to validate our assumptions that:

(1) a voice-based interface would be preferred to a text-based interface as it enables a natural interaction modality [15];
(2) patients would understand that the ECA replaces only non-meaningful work (*benevolence*);
(3) patients would perceive the ECA as respectful of patients' privacy and compliant with regulations because of the high reputation of the clinical hospital (*integrity*);
(4) the embodiment we shaped–a cartoon-like agent wearing an *Anonymous* hospital coat–would contribute to conveying the agents' competence (*ability*, *integrity*);
(5) a referencing conversational style–the agent referring to expert personnel of the Erasmus University Medical Center–would lead patients to perceive the ECA as competent and reliable (*ability*, *integrity*).

To validate our assumptions we conducted a between-subjects Wizard-of-Oz (WoZ) study where one-half of the participants experienced the procedure with the chatbot, and the other half experienced it with the ECA (see Table 3). As a growing body of literature warns about how lack of comparison groups in qualitative research may hinder the way to drawing conclusions about differences and similarities observed in the studies [32], we used the chatbot for assessing the impact of using an embodied agent and natural interaction modalities (voice and body expressivity) rather than text; whether the natural interaction and the appearance would contribute to a positive perception of the agent.

The chatbot was developed using an online web design platform that allows live testing of interface prototypes, which we leveraged for WoZ testing. The appearance of the chatbot was designed to match the aesthetics of the OCHP *Smarter Pregnancy* platform and placed on a screenshot of the real interface to give patients the impression that it was implemented (see Figure 3).

In addition to the comparison between the chatbot and the ECA, we further distinguished the experiential conditions in referencing and non-referencing conversational style. We tested a total of four experimental conditions:

- chatbot referencing (chatbot ref)
- chatbot non-referencing (chatbot no-ref)
- ECA referencing (ECA ref)
- ECA non-referencing (ECA no-ref)

Each of these conditions was tested by at least five (5) patients (see distribution of participants across conditions in Table 2).

4.1 Setup and procedure

As CAs, like other complex interactive systems, promise large interactional benefits but remain a challenging enterprise in terms of development [50], we conducted a Wizard-of-Oz (WoZ) study in which a researcher of the team remotely controlled the CAs. We opted for a WoZ study because it allows participants to envision the intended future interactions [44] and elicits much more complete information compared to other prototyping techniques as it may deliver a close-to-complete specification of the intended system [2, 50]. We set up a monitor in a consult room where the patients would be invited to sit and the session with the CAs would be run. In an adjacent room, we placed the wizard setup consisting of a laptop, an external screen, and a webcam.

In WoZ studies it is desirable that participants believe that they are interacting with a real system, thus they should not be told the truth about the procedure in advance, yet for ethical reasons (i.e., patients' manipulation) researchers should also not lie [2]. We followed Bernsen et al. [2] recommendation to provide vague information that would lead participants to interpret the CAs as if were real. We informed participants that during the session they would be interacting with an experimental conversational system, without any specification regarding the type of agent or details about the connection between these and the OCHP *Smarter Pregnancy* platform. Participants were then informed about the WoZ protocol after the testing when meeting with the researcher for the interview.

While developing a functioning chatbot for our study would have been feasible, we decided to run both conditions (chatbot and ECA) as WoZ to ensure that the same procedure and conversational style would be followed. Leveraging the conversational capabilities of a wizard, however, has also risks and potential limitations. As Breazeal et al. [7] observe in the case of social robotics, relying on the capabilities of the wizard may lead research to bypass errors and technical limitations that would significantly impact the actual development of such systems. To address this issue, WoZ experiments should be designed as rigorous and repeatable procedures,

Table 2: Example sentences for the two different conversational styles and style main characteristics

	Referencing style	Non-referencing style
characteristics	neutral, calm, conveyer of information, like a counsellor	confident, caring, empathic, personal
small talks	none, mainly factual, effective, efficient	confident, caring, empathic, personal
inquiring information	systematic, standardized **example**: "We will move on to the next question, because of our limited time."	based on answers by the patient, let conversation flow **example**: "Could you tell me more about that" "And how does that affect your sleep?"
answering questions	referencing to experts **example**: "According to X from institute it is better to eat 300 grams of vegetables per day"	based on its own knowledge **example**: "It is better to eat 300 grams of vegetables per day"

allowing for a smoother transition toward implementation [44]. Thereafter, we instructed the wizard to follow a scripted series of questions and answers, defined by the researchers performing as the wizard together with the nurse involved in the experimental sessions. The procedure was rehearsed several times with the nurse before being run with patients. The wizard would provide the questions and answers either textually, in the case of the chatbot, or verbally, in the case of the ECA. We limited the possibilities for open-ended conversations to any unexpected questions (e.g., asking for clarifications). If unexpected questions would require specific clinical knowledge, the wizard would respond to contact a nurse or an obstetrician.

The procedure consisted of three main steps:

- *Information.* A nurse and the researcher together welcome the patient (and their partner, if present) to the consult room and inform her (them) about the session with the CA. The participant(s) is also invited to read and sign the Informed Consent Form. After informing the participant(s), the nurse and the researcher leave the room.
- *Interaction.* The participant(s) attend the consult run by the CA. The session lasts about five minutes.
- *Feedback.* The researcher joins back the consult room and conducts a semi-structured interview with the participant(s).

4.2 Participants

A total of 25 participants (N = 25) took part in the study. Of these, 18 were female and 7 were male (F = 18; M = 7). In accordance with ethical approval, the collection of demographic data was not permitted for this study. All female participants were women participating in the periconception care program. Male participants were accompanying their female partners who were interacting directly with the CAs. The sessions were carried out over a period of 9 days with an average of 3 consultations per day. Participants were recruited by inviting women participating in the periconception care program who had to attend their first consult on one of the days when the study was run. We made clear that they could decline the invitation to join the experiment and have a regular consult. A small number of patients refused to participate.

Table 3: Participants' distribution across experimental conditions

Chatbot Ref	Chatbot No-Ref	ECA Ref	ECA No-Ref
P1	P8	P14	P7
P2	P9	P15	P12
P3	P10	P16	P19
P4	P11	P17	P20
P5	P23	P18	P21
P6	P25		P24
P13			
P22			

4.3 Data collection and analysis

We collected feedback from participants through semi-structured interviews. Participants were invited to first share general impressions and beliefs regarding the feasibility of using CAs in healthcare, and whether they felt familiar with the agent they interacted with. Afterwards, participants were invited to reflect on whether they felt the agent was competent, whether they perceived it as a useful resource and whether they perceived risks related to the use of CAs in this kind of care program.

Due to the sensitive context, interviews were not recorded. The researcher conducting the interviews noted participants' responses and subsequently transcribed them digitally. We analyzed and coded the transcripts following deductive thematic analysis and iteratively complemented this with open coding. Building on previous healthcare studies where comparison groups were used (see the review by Lindsay [32]), we first coded all transcripts from both groups and conditions and afterwards, we compared and contrasted similarities and differences between the groups. We coded the transcripts according to the five assumptions listed above, which were grounded on previous literature on the benefits of ECAs in healthcare (i.e., providing natural interaction modalities) and on Mayer's [36] theory of factors affecting trustworthiness(*ability*, *benevolence*, and *integrity*).

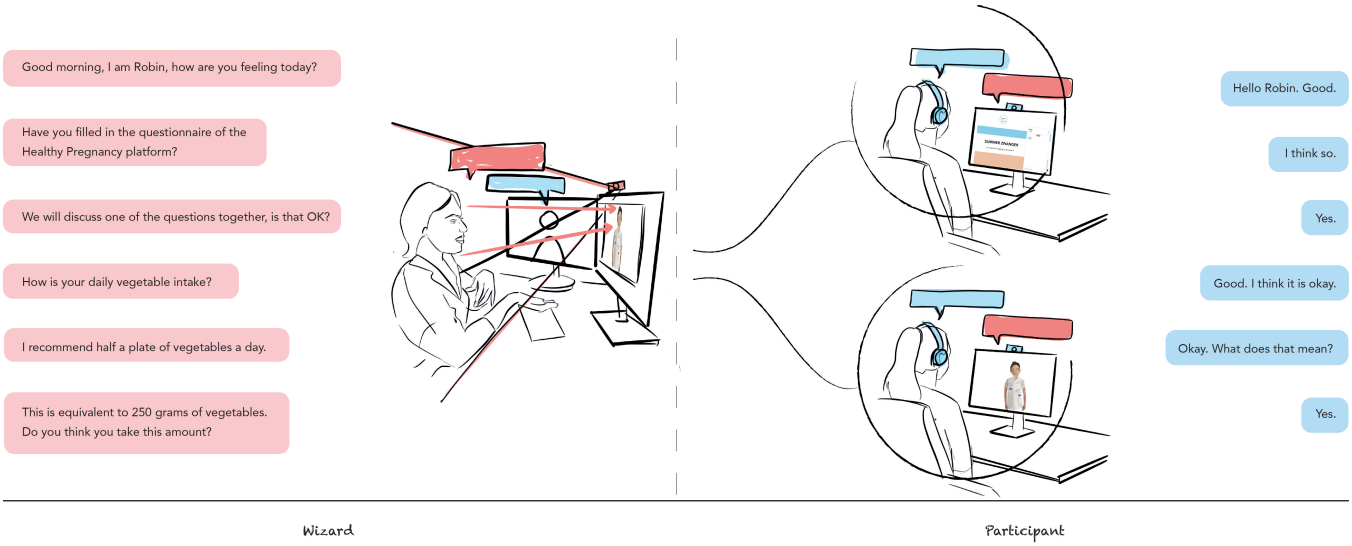

Figure 6: Study set-up with example snippets of conversation between the patient and the ECA or the chatbot

4.4 Ethical approval

The participants of our research were not subjected to any actions or imposed a particular mode of behaviour and thus not considered as a Medical Research Involving Human Subjects Act (WMO) mandatory research. Nevertheless, a description of the protocol for the test was mandatory also for non-WMO mandatory research. We drafted this by following a template from Erasmus University Medical Center, which included a description of the purpose of the study and a list of topics, and submitted it to the Medical Ethics Assessment Committee (METC). Before each study session, participants were provided with a Patient Information Form (PIF) to read and sign, in order to agree to participate in the study. The PIF contained information regarding the study procedure and data treatment so that potential participants could make an informed decision about whether or not to take part in the research.

5 RESULTS

In this section, we illustrate the results of our analysis divided into five sub-sections, each responding to the assumptions listed in section 4.

5.1 Different agents, different mixed feelings

Our assumption that a voice-based interface (ECA) would be preferred to a text-based interface (chatbot) because it enables a natural interaction modality [15], was not confirmed. Both the chatbot and the ECA, in fact, raised almost equally distributed positive and negative comments. About the chatbot, in some positive comments, we noticed appreciation of the interaction modality because of its familiarity, such as *"I think our generation is used to being asked questions like that"* (P3) and *"Nowadays everything has to be via phone and computer"* (P11). Yet, some participants also criticized the very act of typing instead of talking in this context (P10, P9), and other issues, such as difficulty for people to understand the chatbot (P4) or the chatbot providing a superficial conversation (P2). The ECA also received both positive and negative feedback. On the one hand, some participants mentioned that they felt comfortable during the session (P17, P19, P20), and one stated that the ECA *"is somehow better than just a phone call"* (P19). On the other hand, others expressed contrasting views: *"it did not feel comfortable to talk with the robot"* (P16). Based on the results, then, we cannot argue that the ECA would be preferable to a text-based conversational interface. Both agents surfaced a mix of positive and negative aspects. Yet, the motivations underlying participants' perceptions and critiques were often grounded on UX issues that differed between the chatbot and the ECA. We expand on this in section 6.1.

5.2 Participants fear losing human contact

As we designed the ECA as part of the existing care program where various data collection tools are already in use, we assumed that patients would understand that the agent is designed with benevolent intentions, and not to replace meaningful interactions (*benevolence*). Many participants, however, raised concerns regarding the use of the ECA for consults (P7, P14, P16, P24) and led participants to compare it with and express a preference for a one-on-one consult with a nurse (P17, P18, P24), with no difference across the two conversational style conditions. Some participants did emphasize possible gains of having an ECA, such as improving efficiency as you *"don't have to wait for the doctor or nurse but could do this in between"* (P14) and having a better experience than the one you might have via phone call (P19). However, the majority of participants, including some manifesting a positive attitude towards the agent, emphasized how direct interaction with a person is always preferable (P7, P14, P16, P17, P18, P24). Some explained that the interaction with the ECA can work in the case that the consult is a short one (P7), but if the patient needs to discuss more serious concerns would not

rely on this (P14). Other comments suggest that the conversations with the ECA, compared to 1-on-1 conversations with a nurse, are perceived to lack personalization (P24), individually tailored advice (P16), and familiarity (P17). And more broadly, some participants expressed a struggle to see the overall need for such solutions, such as one stating *"in corona time I understand that this could be a part of Erasmus MC. Otherwise, no"* (P16) or another arguing *"I do understand people staying at home to have this conversation, but just do me rather over the phone"* (P21).

5.3 Context may affect integrity concerns

Regarding *integrity*, our hypothesis was that the high reputation of the clinical hospital would be sufficient for participants to perceive the ECA as respectful of their privacy and compliant with regulations. In line with our assumptions, a very limited number of participants' responses during the interviews addressed aspects of *integrity*, with no explicit differences among the various conditions (referencing style vs non-referencing style; chatbot vs ECA). This result was expected because, as we explain in section 3.1, we relied on the hospital's reputation in terms of privacy and on established procedures of the periconception care program for the interaction session. Furthermore, participants signed an informed consent form where they were informed about how data collected during the program is managed. Interestingly, none of the few concerns that emerged regarding integrity were raised by participants who experienced the ECA. For instance, concerns were raised by P9 who argued that *"you do want to know who is behind it (the agent). That there's a little more... security. These days they ask for so much online and they are in your e-mail in no time"*. And P5 mentioned that *"there was not enough questioning to make the conversation deep"*, which is similar to P2's opinion that the agent *"can only give advice and suggestions, because of this I don't really feel the need to change my lifestyle"*. This suggests that the expectations that some participants had on what the consult would be were not satisfied. Although this issue did not emerge from people experiencing the ECA in our study, it may be still important to consider when designing these agents and the protocols for introducing them to patients. To properly assess this aspect, however, further comparative research would be needed (comparing the same ECA in different environments with different reputations).

5.4 Embodiment may be superfluous

Building on previous literature [57] and [28], we assumed that a cartoon-like agent wearing an Erasmus MC hospital coat would be perceived as a non-uncanny yet competent (*ability*, *integrity*). Results, however, showed that this was true only in very few cases. One participant (P12) mentioned that they *"had the idea that it was part of the Erasmus MC*", and another emphasized that the *coat just makes the picture complete"* (P7). Conversely, others expressed aversion towards the ECA, for instance, P14 mentions that the ECA is *"some kind of animation or computer puppet, and it's a little weird to have a conversation with that"*. Nevertheless, most of the participants' responses do not present a strong stand on either the positive or negative side. P15, for instance, explained that *"(the coat) specifically adds anything for me. With other clothing, it would have been the same"*. Relatedly, P21 and P20 both mentioned that they did not pay much attention to the appearance of the ECA, and rather focused on the interaction and what would come next in the conversation. Thereafter, responses regarding the ECA suggest that this type of embodiment may have neither the positive role we devised nor a negative effect. To properly assess this aspect, however, further comparative research would be needed (comparing different appearances for the ECA).

5.5 Referencing undermines competence

We assumed that a referencing conversational style–the agent referring to expert personnel of the Erasmus University Medical Center when providing information and recommendations–would help patients to perceive the ECA as competent and reliable (*ability*, *integrity*). Many comments from participants, indeed, suggest that the ECA was perceived as competent in performing their tasks; but mostly in the condition presenting a non-referencing conversational style. The non-referencing ECA was mostly appreciated by five out of the six participants who experienced it. In fact, despite also mentioning limitations and possible concerns, participants mentioned that the non-referencing ECA works well (P7, P12), it is credible, (P20, P24), it may be preferable to a phone call (P19), and that they can imagine it being used in the future (P12, P20, P24). Only one participant who experienced the non-referencing ECA expressed almost exclusively negative feedback, such as *"I couldn't take it seriously. I found it uncomfortable"* (P21). Regarding the referencing ECA, instead, only one participant (P17) expressed predominantly positive feedback, such as *"it worked very well. [...] She responded well to me. Because of this, she understood me"*, suggesting that the agent is perceived as competent. The rest of the participants, five out of six who experienced the referencing ECA, mostly expressed unease (P16, P14, P15, P18), and the overall impression of talking with an automated system from which you can only get standard answers (P14) and if you want more depth or discussion that deviates from the scripted questions and answers, it doesn't work (P18). Even if none of the participants explicitly related their concerns about the referencing ECA to the conversation style, the significantly different results between the two conditions suggest that the non-referencing ECA is perceived by patients as more credible and competent.

6 DISCUSSION

Our WoZ study with patients confronted us with partially unexpected results. Building on CAs literature [8, 11, 28, 33, 38, 47, 50], we expected that the ECA we developed would be perceived as trustworthy, as it would convey ability, benevolence, and integrity. This, however, was only partially true. Learning from [57] and [28], we assumed that the appearance (a female character wearing an Erasmus MC hospital coat) and conversational style (the agent referring to expert personnel of the Erasmus University Medical Center), would contribute to perceiving the ECA as competent and valuable for patients. Results instead revealed that the appearance of the ECA had a very limited effect on how patients perceived the agent. Even worse, the ECA with the referencing conversational style had the opposite effect than what we expected. Almost all participants who experienced the referencing ECA, expressed unease, and the overall impression of talking with an automated system

unable to engage in discussions that deviate from scripted questions and answers. Lastly, the results also revealed that the ECA raised questions of appropriateness and fostered comparisons between the agents and a nurse.

While, at first sight, these may seem like negative results, we were actually not very surprised and believe the investigation provides interesting insights into designing trustworthy ECAs. First, while only a few instances pointed to a positive influence of the ECA on the experience we also encountered very few responses surfacing feelings of eeriness, which we explicitly tried to prevent through our design. In the few cases where the awkwardness of talking to the ECA was mentioned, this was explicitly discussed along with usability issues of the system. With regards to the design of the referencing conversation style, we were aware that existing literature reported how CAs providing information through a confident and expert style may be preferred over a generalist style [66, 67]. Yet, we built on other examples that show how a CA can be successful and desirable, even if designed as a mediator between a patient and a healthcare professional, such as the work by Kowatsch and colleagues [28]. This specific design choice was grounded on the idea that while a system may be trusted, it may not necessarily be trustworthy [41]. For meaningful reliance and collaboration with artificially intelligent (AI) agents, people must be enabled to build an adequate mental model of the AI agent and its capabilities [9, 59]. Only by understanding what the agent knows or does not know, people can properly calibrate their trust towards the agent [59]. Thereafter, the referencing conversational style is adopted as a strategy to manifest the nature of ECA's knowledge and actions: the first is reliant on the expert knowledge of the healthcare professionals, and the second is limited to scripted questions and answers.

From this perspective, the results are actually encouraging. While full trust and mere reliance on the ECA should be seen as a potentially dangerous sign, patients' capacity to distinguish 'when' to trust the ECA may be considered a sign of a properly calibrated trust. In our results, we see a clear distinction between when patients consider it acceptable to rely on the ECA, e.g., for routine questions and when nothing serious is happening, versus when they would not accept that, such as in the case of problems or difficult situations, in which talking to a nurse would be preferred. Thereafter, *the referencing conversational style could play a crucial role in designing trustworthy ECAs (and CAs in general) as it can discourage the possible over-reliance of patients towards the agent.* One could argue, however, that the referencing conversational style may have a negative impact on the reliance of patience on the agent, beyond the times when this is desirable (i.e. unusual situations). In this regard, while this may become true, such as in the case of prolonged interaction, we believe it would not represent a problem within the use scenario we designed for, as the interaction session would always have a limited duration (15-30 minutes) and recur once every month or every two weeks max. Nevertheless, we identified other usability (UX) issues, often associated with negative feedback towards the ECA, both with the referencing and non-referencing conversational style. UX issues deserve careful consideration, as these could have a significant impact on the agent's acceptability.

6.1 UX design challenges in designing ECAs

The successful design of trustworthy ECAs, and of artificial agents in general, depends heavily on the capacity to build a problem-free operation [28]. While some errors may be tolerated [49], technical issues (such as failures in speech recognition) and the related poor user experience should be limited as these have a negative impact on the intention to use, adaptiveness, usefulness, and trust of people [53]. In our study, we perceived patients' frustration and dissatisfaction, especially in the few moments in which the ECA seemed *"not to pay attention"* (P19), and the sound was lagging behind (P12, P19, P21). Given the Wizard of Oz setup of our study, these UX issues actually originated from errors of the researcher performing as a wizard. Despite the rehearsal sessions, flaws in the performance did occur. This is an inherent possibility in performing Wizard of Oz studies [2], especially if the wizard is not a professional performer. To address this issue, we recommend researchers *include error mitigation strategies in the protocol*, e.g., a sentence justifying why there was a delay in the response, so that the person interacting with the ECA may perceive it as competent despite the error.

The level of fidelity and detail of the prototypes is another important aspect that can play a significant role in the success of a study with ECAs. In our study, we observed a general preference for the chatbot rather than the ECA. This, however, was partially influenced by the level of refinement of the two CAs. In this regard, some participants explicitly mentioned that the ECA feels still in its infancy, whereas the chatbot was often appreciated for presenting a familiar interaction and a smooth operation. As a matter of fact, the two prototypes did have a slightly different level of fidelity, due to the tools used for developing them. While there is a plethora of existing resources and tools for prototyping polished and smoothly functioning chatbot interfaces, the design and prototyping of ECA are still case specific and heavily dependent on the researchers' expertise. In our case, we leveraged an existing example available in the library of the animation tool to simulate the functioning of *Robin*. While practical, this also set some limitations in terms of the level of interactivity and expressivity our ECA could have. Specifically, we only tracked the movement of the head and torso, as well as the facial expressions. We did not animate the arms and hands, which could have been useful to communicate the agent's active listening, showing interest or thinking, which have been shown to have a positive influence on people's trust towards ECAs. Our recommendation, then, would be to *make sure the prototypes under comparison present the same level of fidelity.* While this may often not be possible from a technical point of view, Wizard of Oz techniques can be put in place to simulate that.

7 CONCLUSIONS

In this work, we illustrated the complexity of designing ECAs and the multifaceted nature of trustworthiness when it comes to using these agents for healthcare applications. In our design intervention, we developed an ECA called Robin, through which we explored whether and how an embodiment and conversational style that would make explicit reference to the expert personnel of the Erasmus University Medical Center would be perceived as trustworthy. Reflecting on the results of a study with real patients, where we deployed Robin in the Erasmus University Medical Center and also

compared it with a chatbot, we further unpacked some of the intricacies of designing for trustworthiness as distinct from designing for trust. Specifically, we learned that embodiment and appearance did not significantly influence participants to perceive the agent's expertise, while the referencing conversational style even played a slightly negative influence.

While seemingly negative, the results about the conversational style made us reflect further on the important difference between trust and trustworthiness. As we learned from existing literature, overtrust towards CAs can have dramatic consequences in the healthcare domain (e.g., if they miss out on important signals of life-threatening situations) [30, 39], and sometimes patients do not have an accurate mental model of the actual capabilities of these agents to properly calibrate their trust towards them. Thereafter, we read our participants' critiques of the *referencing ECA*, who would not trust the agent when it comes to difficult situations and serious issues, as a positive sign. Even more so, we believe that the *the referencing conversational style could play a crucial role in designing trustworthy ECAs (and CAs in general) as it can discourage the possible over-reliance of patients towards the agent.* Nevertheless, some participants' critiques pointed to some limitations of our work that should be accounted for. In particular, our study setup surfaced some UX issues that partially influenced negative responses to our ECA, such as belayed responses and a limited level of detailing of the ECA prototype. In this regard, we provide two methodological design recommendations to prevent such UX issues. On the one hand, whether running a study in Wizard of Oz or not, researchers should *include error mitigation strategies in their protocol*, so that participants' trust in the agent's capabilities won't be completely lost in case of errors. On the other hand, it is crucial to *make sure the prototypes under comparison present the same level of fidelity*, so that results would respond to the manipulated features, rather than being biased by the quality of the prototype itself.

Our study presented also methodological limitations that we believe are inherent to conducting research in real clinical environments. The recruiting of participants was challenging at times and some of the people who refused to take part in the study grounded their decision on personal factors like emotional state or disease background. As such, we inadvertently might have introduced a selection bias in our recruitment of participants (e.g., only overall healthy and technologically confident people may be inclined to take part in the interaction). This could be addressed in future research, e.g., by employing randomization in recruitment. Another issue with involving real patients was the lack of control we could exercise over the duration of the sessions. As patients often arrived later or did not show up to the consult, mostly because of forgetting about the appointment, the test sessions were often shortened. In fact, each patient had to attend multiple other consults after the one object of the study. As a result, some patients might have had a poor impression of the agents because of the limited interaction with them. Last, the fact that for ethical reasons each patient had to be welcomed by a nurse before being introduced to the study setup, created confusion about the value of having a CA. Simply put, patients wondered why they had to attend the consult with the CA when the nurse was actually available. While the motivation for and use scenario of CAs for the periconception care program was explained, patients did remain somewhat perplexed.

Despite introducing these challenges to our study, contextualising our work within the specific case of women during the periconceptional and pregnancy period, and running an in-wild study in the Erasmus University Medical Center allowed us to get a rich and situated understanding of both the potentials of ECAs for healthcare, as well as the complexity to design them as genuinely trustworthy agents. The process revealed to us that even more than in other contexts, clinical environments require a continuous collaborative effort among disciplines and stakeholders. From the early to the late stages of any eHealth application, we must develop platforms and provide conditions for continuous feedback loops between patients, clinical personnel, and designers, especially if we are developing AI-powered ECAs.

Recent advancements in generative AI models (GAI), in fact, are showing great potential for revolutionizing eHealth applications and speeding up CAs development. As these may be used to personalize interactions with (E)CAs, GAI have the potential to address many of the challenges presented in CA literature, including our work, such as UX issues related to errors in speech recognition. But it also comes with promises of hyper-personalization which could have a significant impact on user trust and, consequently, controversial implications in terms of trustworthiness. ECAs could be modelled in a way that perfectly fits the needs (e.g., modulating voice speed for hearing impaired patients) but also the preferences of a patient (e.g., the agent presenting the same ethnic traits as the patient). *But what if such preferences turn into yet another mechanism for exacerbating discrepancies because of gender, ethnicity, abilities and more? What if personalization gets in the way of expert recommendations?* With this vision on the horizon, the research community should stay vigilant and tackle the risks that uncritical approaches to AI may have for society at large. As such, this work contributes to building a critical mindset towards CAs, and AI more broadly, by showing the complexity and intricacies of designing for trustworthiness, where designing the most usable and likeable agents may not necessarily be the result we should strive for.

ACKNOWLEDGMENTS

We would like to thank Lorette Paas, nurse at Erasmus MC, who passionately contributed to our project with professional advice and rich insights.

REFERENCES

[1] Alexander M Aroyo, Jan De Bruyne, Orian Dheu, Eduard Fosch-Villaronga, Aleksei Gudkov, Holly Hoch, Steve Jones, Christoph Lutz, Henrik Sætra, Mads Solberg, et al. 2021. Overtrusting robots: Setting a research agenda to mitigate overtrust in automation. *Paladyn, Journal of Behavioral Robotics* 12, 1 (2021), 423–436.

[2] Niels Ole Bernsen, Hans Dybkjær, and Laila Dybkjær. 1994. Wizard of oz prototyping: How and when. *Proc. CCI Working Papers Cognit. Sci./HCI, Roskilde, Denmark* (1994).

[3] Timothy Bickmore, Ha Trinh, Reza Asadi, and Stefan Olafsson. 2018. Safety first: conversational agents for health care. In *Studies in conversational UX design*. Springer, 33–57.

[4] Timothy W Bickmore, Laura M Pfeifer, and Brian W Jack. 2009. Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1265–1274.

[5] Michael Bonfert, Nima Zargham, Florian Saade, Robert Porzel, and Rainer Malaka. 2021. An Evaluation of Visual Embodiment for Voice Assistants on Smart Displays. In *CUI 2021-3rd Conference on Conversational User Interfaces*. 1–11.

[6] Sheryl Brahnam and Antonella De Angeli. 2012. Gender affordances of conversational agents. *Interacting with Computers* 24, 3 (2012), 139–153.

[7] Cynthia Breazeal, Cory D Kidd, Andrea Lockerd Thomaz, Guy Hoffman, and Matt Berlin. 2005. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *2005 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 708–713.

[8] Lorainne Tudor Car, Dhakshenya Ardhithy Dhinagaran, Bhone Myint Kyaw, Tobias Kowatsch, Shafiq Joty, Yin-Leng Theng, Rifat Atun, et al. 2020. Conversational agents in health care: scoping review and conceptual analysis. *Journal of medical Internet research* 22, 8 (2020), e17158.

[9] Luciano Cavalcante Siebert, Maria Luce Lupetti, Evgeni Aizenberg, Niek Beckers, Arkady Zgonnikov, Herman Veluwenkamp, David Abbink, Elisa Giaccardi, Geert-Jan Houben, Catholijn M Jonker, et al. 2022. Meaningful human control: Actionable properties for AI system development. *AI and Ethics* (2022), 1–15.

[10] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, et al. 2019. What makes a good conversation? Challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[11] Enrico Coiera, Baki Kocaballi, John Halamka, and Liliana Laranjo. 2018. The digital scribe. *NPJ digital medicine* 1, 1 (2018), 1–5.

[12] E Coiera, JI Westbrook, and JC Wyatt. 2006. Section 1: Health and Clinical Mangement: The Safety and Quality of Decision Support Systems. *Yearbook of medical informatics* 15, 01 (2006), 20–25.

[13] Benjamin R Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. " What can i help you with?" infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–12.

[14] Andrew J Cowell and Kay M Stanney. 2003. Embodiment and interaction guidelines for designing credible, trustworthy embodied conversational agents. In *International workshop on intelligent virtual agents*. Springer, 301–309.

[15] Samuel Rhys Cox and Wei Tsang Ooi. 2022. Does Chatbot Language Formality Affect Users' Self-Disclosure?. In *Proceedings of the 4th Conference on Conversational User Interfaces*. 1–13.

[16] Nigel Crisp and Lincoln Chen. 2014. Global supply of health professionals. *New England Journal of Medicine* 370, 10 (2014), 950–957.

[17] Ewart J De Visser, Marieke MM Peeters, Malte F Jung, Spencer Kohn, Tyler H Shaw, Richard Pak, and Mark A Neerincx. 2020. Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics* 12, 2 (2020), 459–478.

[18] Ahmed Fadhil, Yunlong Wang, and Harald Reiterer. 2019. Assistive conversational agent for health coaching: a validation study. *Methods of information in medicine* 58, 01 (2019), 009–023.

[19] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association* 19, 1 (2012), 121–127.

[20] Batoul Hojeij, Sam Schoenmakers, Sten Willemsen, Lenie van Rossem, Andras Dinnyes, Melek Rousian, and Regine PM Steegers-Theunissen. 2023. The Effect of an eHealth Coaching Program (Smarter Pregnancy) on Attitudes and Practices Toward Periconception Lifestyle Behaviors in Women Attempting Pregnancy: Prospective Study. *Journal of Medical Internet Research* 25 (2023), e39321.

[21] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. " Because AI is 100% right and safe": User Attitudes and Sources of AI Authority in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.

[22] Rabia Khan and Antonella De Angeli. 2009. The attractiveness stereotype in the evaluation of embodied conversational agents. In *IFIP Conference on Human-Computer Interaction*. Springer, 85–97.

[23] Rabia Fatima Khan and Alistair Sutcliffe. 2014. Attractive agents are more persuasive. *International Journal of Human-Computer Interaction* 30, 2 (2014), 142–150.

[24] Ahmet Baki Kocaballi, Shlomo Berkovsky, Juan C Quiroz, Liliana Laranjo, Huong Ly Tong, Dana Rezazadegan, Agustina Briatore, and Enrico Coiera. 2019. The personalization of conversational agents in health care: systematic review. *Journal of medical Internet research* 21, 11 (2019), e15360.

[25] Rafal Kocielnik, Raina Langevin, James S George, Shota Akenaga, Amelia Wang, Darwin P Jones, Alexander Argyle, Callan Fockele, Layla Anderson, Dennis T Hsieh, et al. 2021. Can I Talk to You about Your Social Needs? Understanding Preference for Conversational User Interface in Health. In *CUI 2021-3rd Conference on Conversational User Interfaces*. 1–10.

[26] Bing Cai Kok and Harold Soh. 2020. Trust in robots: Challenges and opportunities. *Current Robotics Reports* 1, 4 (2020), 297–309.

[27] Ross Koppel, Joshua P Metlay, Abigail Cohen, Brian Abaluck, A Russell Localio, Stephen E Kimmel, and Brian L Strom. 2005. Role of computerized physician order entry systems in facilitating medication errors. *Jama* 293, 10 (2005), 1197–1203.

[28] Tobias Kowatsch, Theresa Schachner, Samira Harperink, Filipe Barata, Ullrich Dittler, Grace Xiao, Catherine Stanger, Florian v Wangenheim, Elgar Fleisch, Helmut Oswald, et al. 2021. Conversational agents as mediating social actors in chronic disease management involving health care professionals, patients, and family members: multisite single-arm feasibility study. *Journal of medical Internet research* 23, 2 (2021), e25060.

[29] Raina Langevin, Ross J Lordon, Thi Avrahami, Benjamin R Cowan, Tad Hirsch, and Gary Hsieh. 2021. Heuristic evaluation of conversational agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[30] Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie YS Lau, et al. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association* 25, 9 (2018), 1248–1258.

[31] Jingyi Li, Michelle X Zhou, Huahai Yang, and Gloria Mark. 2017. Confiding in and listening to virtual agents: The effect of personality. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. 275–286.

[32] Sally Lindsay. 2019. Five approaches to qualitative comparison groups in health research: a scoping review. *Qualitative health research* 29, 3 (2019), 455–468.

[33] Christine Lisetti, Ugan Yasavur, Claudia De Leon, Reza Amini, Ubbo Visser, and Naphtali Rishe. 2012. Building an on-demand avatar-based health intervention for behavior change. In *Twenty-Fifth International FLAIRS Conference*.

[34] David D Luxton. 2020. Ethical implications of conversational agents in global public health. *Bulletin of the World Health Organization* 98, 4 (2020), 285.

[35] Raju Maharjan, Darius Adam Rohani, Per Bækgaard, Jakob Bardram, and Kevin Doherty. 2021. Can We Talk? Design Implications for the Questionnaire-Driven Self-Report of Health and Wellbeing via Conversational Agent. In *Proceedings of the 3rd Conference on Conversational User Interfaces* (Bilbao (online), Spain) *(CUI '21)*. Association for Computing Machinery, New York, NY, USA, Article 5, 11 pages. https://doi.org/10.1145/3469595.3469600

[36] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.

[37] Jean-Arthur Micoulaud-Franchi, Patricia Sagaspe, Etienne de Sevin, Stéphanie Bioulac, Alain Sauteraud, and Pierre Philip. 2016. Acceptability of embodied conversational agent in a health care context. In *International conference on intelligent virtual agents*. Springer, 416–419.

[38] Madison Milne-Ives, Caroline de Cock, Ernest Lim, Melissa Harper Shehadeh, Nick de Pennington, Guy Mole, Eduardo Normando, Edward Meinert, et al. 2020. The effectiveness of artificial intelligence conversational agents in health care: systematic review. *Journal of medical Internet research* 22, 10 (2020), e20346.

[39] Adam S Miner, Arnold Milstein, Stephen Schueller, Roshini Hegde, Christina Mangurian, and Eleni Linos. 2016. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA internal medicine* 176, 5 (2016), 619–625.

[40] David C Mohr, Justin K Benzer, and Gary J Young. 2013. Provider workload and quality of care in primary care settings: moderating role of relational climate. *Medical care* (2013), 108–114.

[41] Onora O'Neill. 2018. Linking trust to trustworthiness. *International Journal of Philosophical Studies* 26, 2 (2018), 293–300.

[42] Alisha Pradhan and Amanda Lazar. 2021. Hey Google, Do You Have a Personality? Designing Personality and Personas for Conversational Agents. In *CUI 2021-3rd Conference on Conversational User Interfaces*. 1–4.

[43] Minjin Rheu, Ji Youn Shin, Wei Peng, and Jina Huh-Yoo. 2021. Systematic review: Trust-building factors and implications for conversational agent design. *International Journal of Human–Computer Interaction* 37, 1 (2021), 81–96.

[44] Laurel D Riek. 2012. Wizard of oz studies in hri: a systematic review and new reporting guidelines. *Journal of Human-Robot Interaction* 1, 1 (2012), 119–136.

[45] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M Howard, and Alan R Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*. IEEE, 101–108.

[46] Melek Rousian, Sam Schoenmakers, Alex J Eggink, Dionne V Gootjes, Anton HJ Koning, Maria PH Koster, Annemarie GMGJ Mulders, Esther B Baart, Irwin KM Reiss, Joop SE Laven, et al. 2021. Cohort profile update: the Rotterdam Periconceptional Cohort and embryonic and fetal measurements using 3D ultrasound and virtual reality techniques. *International Journal of Epidemiology* 50, 5 (2021), 1426–1427l.

[47] Elayne Ruane, Abeba Birhane, and Anthony Ventresque. 2019. Conversational AI: Social and Ethical Considerations.. In *AICS*. 104–115.

[48] Rahime Belen Sağlam and Jason RC Nurse. 2020. Is your chatbot GDPR compliant? Open issues in agent design. In *Proceedings of the 2nd Conference on Conversational User Interfaces*. 1–3.

[49] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*. 141–148.

[50] Jan Smeddinck, Kamila Wajda, Adeel Naveed, Leen Touma, Yuting Chen, Muhammad Abu Hasan, Muhammad Waqas Latif, and Robert Porzel. 2010. QuickWoZ: a multi-purpose wizard-of-oz framework for experiments with embodied conversational agents. In *Proceedings of the 15th international conference on Intelligent user interfaces*. 427–428.

[51] Brendan Spillane, Emer Gilmartin, Christian Saam, and Vincent Wade. 2019. Issues relating to trust in care agents for the elderly. In *Proceedings of the 1st International Conference on Conversational User Interfaces*. 1–3.

[52] Pieter Jan Stappers and Elisa Giaccardi. 2017. Research through design. In *The encyclopedia of human-computer interaction*. The Interaction Design Foundation, 1–94.

[53] Vera Stara, Benjamin Vera, Daniel Bolliger, Lorena Rossi, Elisa Felici, Mirko Di Rosa, Michiel de Jong, Susy Paolini, et al. 2021. Usability and acceptance of the embodied conversational agent Anne by people with dementia and their caregivers: exploratory study in home environment settings. *JMIR mHealth and uHealth* 9, 6 (2021), e25891.

[54] Regine PM Steegers-Theunissen, John Twigt, Valerie Pestinger, and Kevin D Sinclair. 2013. The periconceptional period, reproduction and long-term health of offspring: the importance of one-carbon metabolism. *Human reproduction update* 19, 6 (2013), 640–655.

[55] Yolande Strengers and Jenny Kennedy. 2021. *The smart wife: Why Siri, Alexa, and other smart home devices need a feminist reboot.*

[56] Silke ter Stal, Lean Leonie Kramer, Monique Tabak, Harm op den Akker, and Hermie Hermens. 2020. Design features of embodied conversational agents in eHealth: a literature review. *International Journal of Human-Computer Studies* 138 (2020), 102409.

[57] Silke ter Stal, Monique Tabak, Harm op den Akker, Tessa Beinema, and Hermie Hermens. 2020. Who do you prefer? The effect of age, gender and role on users' first impressions of embodied conversational agents in eHealth. *International Journal of Human–Computer Interaction* 36, 9 (2020), 881–892.

[58] Markus Thaler, Stephan Schlögl, and Aleksander Groth. 2020. Agent vs. Avatar: Comparing embodied conversational agents concerning characteristics of the uncanny valley. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*. IEEE, 1–6.

[59] Richard Tomsett, Alun Preece, Dave Braines, Federico Cerutti, Supriyo Chakraborty, Mani Srivastava, Gavin Pearson, and Lance Kaplan. 2020. Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns* 1, 4 (2020), 100049.

[60] Aditya Nrusimha Vaidyam, Hannah Wisniewski, John David Halamka, Matcheri S Kashavan, and John Blake Torous. 2019. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Canadian Journal of Psychiatry* 64, 7 (2019), 456–464.

[61] Pieter Van den Hombergh, Beat Künzi, Glyn Elwyn, Jan van Doremalen, Reinier Akkermans, Richard Grol, and Michel Wensing. 2009. High workload and job stress are associated with lower practice performance in general practice: an observational study in 239 general practices in the Netherlands. *BMC Health Services Research* 9, 1 (2009), 1–8.

[62] Alastair van Heerden, Xolani Ntinga, and Khanya Vilakazi. 2017. The potential of conversational agents to provide a rapid HIV counseling and testing services. In *2017 international conference on the frontiers and advances in data science (FADS)*. IEEE, 80–85.

[63] Anke Versluis, Kyma Schnoor, Niels H Chavannes, Esther PWA Talboom-Kamp, et al. 2022. Direct Access for Patients to Diagnostic Testing and Results Using eHealth: Systematic Review on eHealth and Diagnostics. *Journal of Medical Internet Research* 24, 1 (2022), e29303.

[64] Sarah Theres Völkel, Samantha Meindl, and Heinrich Hussmann. 2021. Manipulating and evaluating levels of personality perceptions of voice assistants through enactment-based dialogue design. In *CUI 2021-3rd Conference on Conversational User Interfaces*. 1–12.

[65] Catharine Wang, Timothy Bickmore, Deborah J Bowen, Tricia Norkunas, MaryAnn Campion, Howard Cabral, Michael Winter, and Michael Paasche-Orlow. 2015. Acceptability and feasibility of a virtual counselor (VICKY) to collect family health histories. *Genetics in Medicine* 17, 10 (2015), 822–830.

[66] Carolin Wienrich, Clemens Reitelbach, and Astrid Carolus. 2021. The trustworthiness of voice assistants in the context of healthcare investigating the effect of perceived expertise on the trustworthiness of voice assistants, providers, data receivers, and automatic speech recognition. *Frontiers in Computer Science* 3 (2021), 685250.

[67] Runting Zhong and Mengyao Ma. 2022. Effects of communication style, anthropomorphic setting and individual differences on older adults using voice assistants in a health context. *BMC geriatrics* 22, 1 (2022), 1–13.