frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 28 June 2023
DOI 10.3389/fcomp.2023.1152532

OPEN ACCESS

EDITED BY
Laurens Rook,
Delft University of Technology, Netherlands

REVIEWED BY
Jan Van Dalen,
Erasmus University Rotterdam, Netherlands
Francesco Setti,
University of Verona, Italy

*CORRESPONDENCE
Mengyu Zhong
✉ mengyu.zhong@it.uu.se

†These authors have contributed equally to this work and share first authorship

RECEIVED 27 January 2023
ACCEPTED 30 May 2023
PUBLISHED 28 June 2023

CITATION
Zhong M, Fraile M, Castellano G and Winkle K (2023) A case study in designing trustworthy interactions: implications for socially assistive robotics. *Front. Comput. Sci.* 5:1152532. doi: 10.3389/fcomp.2023.1152532

A case study in designing trustworthy interactions: implications for socially assistive robotics

Mengyu Zhong[1,2*†], Marc Fraile[1†], Ginevra Castellano[1] and Katie Winkle[1]

[1]Department of Information Technology, Uppsala University, Uppsala, Sweden, [2]Women's Mental Health During the Reproductive Lifespan—WOMHER Centre, Uppsala University, Uppsala, Sweden

This work is a case study in applying recent, high-level ethical guidelines, specifically concerning transparency and anthropomorphisation, to Human-Robot Interaction (HRI) design practice for a real-world Socially Assistive Robot (SAR) application. We utilize an online study to investigate how the perception and efficacy of SARs might be influenced by this design practice, examining how robot utterances and display manipulations influence perceptions of the robot and the medical recommendations it gives. Our results suggest that applying transparency policies can improve the SAR's effectiveness without harming its perceived anthropomorphism. However, our objective measures suggest participant understanding of the robot's decision-making process remained low across conditions. Furthermore, verbal anthropomorphisation does not seem to affect the perception or efficacy of the robot.

KEYWORDS

human-robot interaction, trustworthy HRI, transparent robots, anthropomorphic robots, ethics, mental healthcare, explainable AI

1. Introduction

An increasing amount of work within Human-Robot Interaction (HRI) is concerned with the design and development of Socially Assistive Robots (SARs) for applications in healthcare (Kyrarini et al., 2021). Many user studies look to examine the impact of particular design choices in this context with regards to robot behavior (Agrigoroaie and Tapus, 2016) or appearance (Cresswell et al., 2018). Unsurprisingly, given the human-centered nature of healthcare as a target application, and a tendency to design robots which aim to emulate human-human interaction cues, SARs are often quite anthropomorphic in their design. A number of scholarly works critique such anthropomorphisation (Wilks, 2010; Yogeeswaran et al., 2016; Danaher, 2020), and an increasing number of ethical guidelines for robot design/development similarly call for careful consideration, on the part of robot designers and developers, as to if/how they responsibly leverage anthropomorphism while also ensuring users understand the actual capabilities and limitations of their systems. Some such guidelines call broadly for the avoidance of "*unnecessary*" anthropomorphisation, and ensuring "*transparency of [the robot's] robotic nature*" (BSI, 2016), raising the immediate questions: when indeed may anthropomorphism be (un)necessary? and how can we ensure transparency? Roesler et al. (2021) performed a meta-analysis suggesting that anthropomorphism generally has positive effects on human-related outcomes, including likeability, intelligence, trust and acceptance. Winkle et al. (2021) also suggested that anthropomorphic behavior may increase the efficacy of SARs.

With the prevalence of Artificial Intelligence (AI) in current robotics and the fact that robots commonly require data to function, the ethical guidelines for trustworthy AI and legislation for data protection also become applicable (Felzmann et al., 2019). Transparency is a recurrent theme in such guidelines. However, similar to anthropomorphism, the effects of transparency can vary across scenarios. Nesset et al. (2021) investigated dialogue-based transparent manipulations and suggested transparent design could lead to better-informed decisions on health. While Straten et al. (2020) raised concerns that transparency could impair the perceived anthropomorphism and trust in robots, leading to reduced relationship formation in a child-robot interaction. As such, it seems timely and valuable to examine:

1. How might robot developers design SAR behaviors and user interactions which deliver on current ethical guidance?
2. What might the implications of such design choices be on *efficacy*, i.e., how well the robot achieves the desired goal of deployment, which includes e.g., being acceptable to users?

We explore these questions by providing a case study within the context of Perinatal Depression (PND) screening. Based on recent work studying automatic detection of PND in late pregnancy and the application of Explainable Artificial Intelligence (XAI) techniques (Lundberg and Lee, 2017; Zhong et al., 2022), we consider a robot-patient interaction, in which a SAR automatically evaluates if the patient is at risk of PND. Addressing question one, we produce design speculations on what lower vs. higher levels of *designed anthropomorphism* could "look like" in this context, in a fixed, commercially available and commonly used robot platform, while respecting laws and regulations. We do the same with *designed transparency*, speculating on how data regulations and XAI techniques might be integrated into SAR interactions. We further investigate whether our more transparent design really supports increased user understanding of the system, as called for by the ethical guidelines (HLEG, 2020). Addressing question two, we conduct an online, video-based study to examine how each version of our SAR is perceived by a relevant participant pool: women with children. Specifically, we examine the impacts of our design manipulations on participants' perceptions of likability, anthropomorphism, and trust in our SAR, representing typical evaluation measures commonly utilized within HRI user studies. We further investigate the extent to which participants (i) seemingly understood the SAR's function, (ii) subjectively perceived the SAR as deceptive or clear and appropriate, and (iii) would likely act upon medical advice given by the SAR.

The video-based nature of our study puts participants in an observing role with minimal "real world" risk. Early work in HRI advocated for the value of such approaches during prototyping, testing and developing HRI scenarios, finding high agreement between live vs. online trials (Woods et al., 2006), and, even before COVID-19 forced many studies online (Feil-Seifer et al., 2020), video-based stimulus have been used to investigate perceptions of robots within HRI research (Blow et al., 2006; Cramer et al., 2009; Lee et al., 2011; Strait et al., 2015; Sanders et al., 2019; Kim et al., 2020; Kwon et al., 2020), particularly when investigating high-impact, ethically and/or morally fraught HRI scenarios (Rosenthal-von der Pütten et al., 2013; Jackson and Williams, 2019; Winkle et al., 2021). In line with these previous works, our work is intended to provide initial insights on currently under-explored practicalities and implications of designing for trustworthy HRI in a highly sensitive-but-current SAR application.

In this work we do not consider the physical embodiment of the robot, but rather variations in behavior design for one specific robot platform—a realistic remit for many HRI designers who work with (a limited number of) commercial platforms. We use a Pepper robot for all manipulations, since its CE marking[1] makes it a realistic option for deployment in-the-wild. We further assume that the robot needs to behave lawfully and avoid actively deceiving users in all experimental conditions. Therefore, we limit *designed anthropomorphism* to conversational manipulations that follow existing guidelines for ethical and trustworthy design. In the *low anthropomorphism* condition, the robot reflects a strict interpretation of the ethical guidelines, never referring to itself as an independent agent, avoiding expressions that would imply it has thoughts or emotions, and attributing agency over its actions to the medical team and the interaction designers. In the *high anthropomorphism* condition, the robot uses more "natural" language and represents itself as a member of the medical team that actively contributes to the decision-making process.

Similarly, we seek to represent *designed transparency* in a realistic manner that can be applied to in-the-wild interactions today. We combine two approaches: (i) disclosure and clarification of data collection and usage in one hand and (ii) explanation of algorithmic processes in the other. In the *low transparency* condition, the robot provides only minimal details about which data is collected and how it is used, as might be required for informed consent, but does not explain how it reaches conclusions. In the *high transparency* condition, the robot is more explicit about data usage. It also uses its display capabilities, combined with real-world XAI techniques to explain certain why and how certain decisions have been made.

The results indicate that our transparent manipulations significantly affected the persuasiveness and perceived clearness of the robot, without influencing its perceived anthropomorphism or perceived ethical risk. However, this did not translate clearly into the pattern of measured participant understanding of the system. When ensuring enough anthropomorphism for smooth interaction, modifying anthropomorphism had no significant effects on the perception nor efficacy of the robot.

2. Related work

2.1. Designing trustworthy human-robot interactions

As AI-driven robots become more prevalent in everyday and professional contexts, where they directly interact with human users, trustworthiness becomes an increasingly important requirement in their design (Kraus et al., 2022). The European Commission's *Ethics guidelines for trustworthy AI* (HLEG, 2019) define three key characteristics to enable trust in an automated system: *lawful*, *ethical*, and *robust*. Many other guidelines,

1 https://www.aldebaran.com/sites/default/files/inline-files/declaration-of-conformity-pepper-1.8.pdf

principles, and legislation for AI and robots also impact the design of trustworthy robots. Some examples include UNICEF's *Policy guidance on AI for children* (UNICEF, 2021), the European Union's *General Data Protection Regulation* (GDPR, 2018), and British Standard BS8611—*guide to the ethical design and application of robots and robotic systems* (BSI, 2016). However, when it comes to real-life scenarios, clear and mature instructions for the practical implementation of trustworthy HRI are still not in place. Very recently, Kraus et al. (2022) attempted to draw a more concrete and applicable checklist for trustworthy HRI. Zicari et al. (2021a) focused on the *assessment* of trustworthy AI by proposing a methodological framework, Z-Inspection. Later the same year, Zicari et al. (2021b) demonstrated this approach in a case study involving AI in healthcare, showing their interpretation of the European Commission's trustworthy AI guidelines. However, to the best of our knowledge, there is no such evaluation of trustworthiness available for HRI.

These high-level guidelines are often designed with a wider context in mind, and do not necessarily translate well to HRI. Some design practices implicate that the guidelines are not ideal for the demanded functions of a social robot. Lemaignan et al. (2021, 2022) applied UNICEF's Policy Guidance on AI for Children (UNICEF, 2021) when designing Pepper robots for autistic children. They appreciated the solid ethical framing, but questioned whether such guidance would be appropriate and applicable to embodied AI systems like SARs.

Winkle et al. (2021) evaluated the ethical risk of anthropomorphism and deception, which is identified in BS8611 by suggesting to avoid *"unnecessary anthropomorphization"* and *"deception due to the behavior and/or appearance of the robot"*, and to ensure *"transparency of its robotic nature"* [page 3 of BS6811 (BSI, 2016)]. The study indicated that the kind of conversation-based anthropomorphism we experiment with in this work is likely important for overall SAR function and represents low ethical risk to users, challenging the anti-anthropomorphisation recommendations. Nevertheless, considering the non-human *"robotic nature"* of the system needs to be transparent, this result causes an interesting practical dilemma: how can we ensure transparency of the system's robotic nature without harming the essential anthropomorphism? Or we can question one step further: how does transparency affect the *perception* and *efficacy* of SARs?

2.2. Transparency and anthropomorphism for trustworthy HRI

2.2.1. Meanings of transparency

Due to the inherent ambiguity of the word, *transparency* has been interpreted differently by various guidelines. The Engineering and Physical Sciences Research Council's (EPSRC) Principles of Robotics, which was first published online in 2011, include a Principle of Transparency that requires a robot's machine nature to be transparent (Boden et al., 2011, 2017). The BS8611 poses the same attitude: *"ensure transparency of its robotic nature"*. In these documents, transparency is implemented as *disclosure of the machine-like nature of the robot*. The EPSRC Principles further state: *"Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent."* (Boden et al., 2017). The normative assertion makes the impression that non-transparency and deception caused are generally unacceptable, as they would lead to exploitation. In fact, the effect of transparency in real-world applications is more complex than that, varying with the application and purpose of the robot (Bryson, 2017; Wortham and Theodorou, 2017).

A separate meaning emerges from legislation protecting the use of personal data. In GDPR (2018), transparency is implemented as *disclosure of the collection and processing of personal data*. This creates new expectations of robot transparency, as robots often function with data. Some general principles are provided in recital 39: *"It should be transparent to natural persons that personal data concerning them are collected, used, consulted or otherwise processed and to what extent the personal data are or will be processed..."*, and recital 58: *"The principle of transparency requires that any information addressed to the public or to the data subject be concise, easily accessible and easy to understand, and that clear and plain language and, additionally, where appropriate, visualization be used...This is of particular relevance in situations where the proliferation of actors and the technological complexity of practice make it difficult for the data subject to know and understand whether, by whom and for what purpose personal data relating to him or her are being collected..."*

A final meaning of transparency is listed in the Assessment List for Trustworthy Artificial Intelligence (HLEG, 2020). Specifically, under the subheading of Transparency, the questions posed include: *"Did you assess to what extent the decisions and hence the outcome made by the AI system can be understood?"*, *"Did you communicate to (end-)users—through a disclaimer or any other means—that they are interacting with an AI system and not with another human?"*, and *"Did you clearly communicate characteristics, limitations and potential shortcomings of the AI system?"* The second question follows the same ethos as the ESPRC Principles and BS8611, while the first and third questions focus on the *explainability* of the system, as well as promoting user awareness of its limitations. We can understand this implementation of transparency as *allowing the end-user to correctly calibrate their trust in the system*.

In the field of robotics, however, the competing notions of transparency are still under-explored. Interpreting these high-level expectations and applying the transparency principles is an open challenge (Felzmann et al., 2019).

2.2.2. Transparency in practice

Empirical studies have applied the transparency principles in robots through both verbal and non-verbal cues, showing various effects on outcomes, including trust, utility, and robustness (Wortham and Theodorou, 2017). Wang et al. (2021) designed three levels of transparency in a robotic driving assistant, by conveying different amounts of information in the auditory and visual communication channels. Their results showed that the same transparency level leads to different outcomes for different tasks, even in the same driving context. Other studies suggest that language-based explanation has positive effects on trust (Wang

et al., 2015; Nesset et al., 2021). Nesset et al. (2021) found transparency is beneficial for users to evaluate the competence of the robot and calibrate trust. Although we could identify more use-cases (e.g., companion robots) in which transparency could be at odds with utility, the relationship between transparency, trust, and utility still needs to be further explored (Wortham and Theodorou, 2017).

2.2.3. Expectations on anthropomorphism

In the realm of HRI, employing anthropomorphism is one of the most prevalent design strategies to enhance people's acceptance of social robots (Fink, 2012; Dörrenbächer et al., 2020). Roesler et al. (2021) suggest that task-relevant anthropomorphic features can increase task performance. Winkle et al. (2021) also claims that anthropomorphism is important for the robot function in SARs. Yet, controversially, anthropomorphism is discouraged in the realm of robot ethics. In Section 2.1 we have highlighted the EPSRC Principles of Robotics (Boden et al., 2011) and British Standard BS8611 (BSI, 2016). Similarly, the trustworthy and acceptable HRI checklist (TA-HRI) (Kraus et al., 2022) contains the following item: *"The robot uses intuitive mechanisms of interpersonal communication appropriately (without excessive anthropomorphising or an inappropriate degree of attachment)"*.

2.2.4. Anthropomorphism in practice

anthropomorphic design can be implemented by moderators such as appearance, movement, interaction/communication, and the context in which the robot is deployed and introduced to users (DiSalvo et al., 2002, 2004; Onnasch and Roesler, 2021). Roesler et al. (2021) note that, depending on the moderators used and other contextual factors, anthropomorphic design is not always beneficial and can lead to diverse and unintended outcomes. However, their meta-analysis found overall positive effects on human-related outcomes, including perceived likeability and intelligence, trust and acceptance toward robots, activation, pleasure, and social behavior. The anthropomorphic communicational features used in Winkle et al. (2021)'s study have also shown significant positive effects across studies (Roesler et al., 2021). Wang et al. (2021) told a different story that for driving context, anthropomorphic visual and audio communication can reduce usability as the workload increases.

2.3. Appropriate trust in robots

The word "trust" gives a positive common feeling that it should always be an important and good thing. Some argue that it is a critical element of human–robot relationships. People's undertrust in robots can curb their willingness to accept robot-produced information or follow a robot's suggestions, thus limiting the potential benefit of robotic systems and affecting the efficacy of robots (Hancock et al., 2011; Schaefer et al., 2016). Whereas, overtrusting can be equally problematic and engender ethical issues like overinvesting in robots and deceptions, which are issues the ethical guidelines intended to address by reducing anthropomorphism. Instead of going too extreme, what is more important might be to enable users to correctly calibrate their level of trust appropriately and avoid over- or under-usage (Dzindolet et al., 2003; Abbass et al., 2018; Hancock et al., 2021), and lead to trustworthy design.

Although transparency (Lewis et al., 2018) and anthropomorphism (Natarajan and Gombolay, 2020; Roesler et al., 2021) are often considered boosters of trust (Hancock et al., 2021), just like "trust", these two ambiguous words can mean different things to different people in different contexts and are not necessarily beneficial for trustworthiness or efficacy (Weller, 2017; Wortham and Theodorou, 2017).

3. Research questions

There is tension in the existing literature between the potential advantages of a highly anthropomorphic robot, and current recommendations to obtain trustworthiness by maximizing transparency and avoiding deception. While this discussion tends to encompass anthropomorphism as a form of deception, these two qualities might be separable. Given the design space afforded by common interaction modalities for a social robot (speech, movement, display) and previous literature, we explore separate manipulation of the *designed anthropomorphism* level (low vs. high) and the *designed transparency* level (low vs. high), with the goal to design a *trustworthy* and *effective* assistant in the healthcare domain. This leads us to evaluate our design through the following research questions:

- **RQ1** How do different levels of *designed anthropomorphism* (low vs. high) and *designed transparency* (low vs. high) affect the *perception* of a social robot as a healthcare assistant screening for PND?
  - **RQ 1.1** How do the experimental conditions affect responses to *standardized questionnaire scales*?
  - **RQ 1.2** How do the experimental conditions affect *scenario-specific measures* related to ethical risk?
- **RQ2** How do different levels of *designed anthropomorphism* (low vs. high) and *designed transparency* (low vs. high) affect the *efficacy* of a social robot as a healthcare assistant screening for PND?
  - **RQ 2.1** How do the experimental conditions affect *agreement* with the robot's recommendations?
  - **RQ 2.2** How do the experimental conditions affect *understanding* of the robot's decision making process?

4. Methods

We constructed an online, video-based user study in which participants watched a recorded interaction between a Pepper robot and an actress playing the role of Mary, a pregnant woman. In the recording, Pepper screens the woman for PND. The study followed a 2×2 between-subjects design, corresponding to the *designed anthropomorphism* level (low vs. high) and the

designed transparency level (low vs. high). These manipulations represent our exploration of the conversation design space, based on the existing literature and ethical guidelines presented under Section 2. Each participant was randomly assigned to one of four versions of the recording, and answered a questionnaire based on their observations.

4.1. Scenario

Our scenario is grounded in previous studies exploring PND prediction through mobile phone data (Zhong et al., 2022), and social robotics in PND (Zhong et al., 2021; Tanqueray et al., 2022). Ultimately, the purpose of the robot is to encourage those predicted to be most at risk of PND to accept a follow-up appointment with a specialist clinician. To evaluate PND risk, we consider two real-world tools: the Edinburgh Postnatal Depression Scale (EPDS), and the Mom2B mobile app (Bilal et al., 2022).

The EPDS (Cox et al., 1987) is a 10-item questionnaire that is widely used as a screening tool for PND in Sweden. Swedish guidelines indicate that the questionnaire should be filled in by the patient, and a health professional should follow up by discussing the results with the patient.[2] The Mom2B phone app (Bilal et al., 2022) collects data monitoring the user's depression-related symptoms throughout their pregnancy, and up to one year postpartum. The data collected includes e.g., sleeping hours, number of steps walked, and survey answers collected via the app. This data can be used to predict the future occurrence of PND. For instance, in late pregnancy (Zhong et al., 2022), or after childbirth (Bilal et al., 2022).

In the fictional scenario presented to the participants, a Pepper robot has been deployed in a health center to assist in screening for PND. Mary, a pregnant woman, visits the center for her scheduled screening, and has a short conversation with Pepper to assess her mental wellbeing. The robot asks for permission to access the patient's app data (from Mom2B or a similar app), which she accepts. Pepper then uses this data to perform a preliminary assessment, determining that the patient is at risk of developing PND. The robot complements its explanation of the assessment with an informational graphic displayed on its screen. It asks the patient to fill in a questionnaire and, after the patient agrees, it shows the EPDS in its built-in screen. Once the patient finishes responding, the robot asks Mary to verbally discuss and explain some of her answers in an open-ended discussion (akin to the real-world delivery of EPDS), noting that this will also be used to assess the risk of PND. Finally, Pepper concludes that the patient is at high risk of developing PND and suggests scheduling a follow-up visit with a doctor. The patient refuses this suggestion.

4.2. Video stimulus

All four video stimuli used in the study were filmed in the same static set-up, as shown in Figure 1: Pepper is centered in the frame as the most important subject, with its front side fully visible. The actress (a member of the research team) occupies the space to the right, and is only seen from the back. Whenever either character speaks, color-coded and labeled subtitles are overlaid on the bottom of the frame (see Figure 1A). The left side only contains background furniture, and is used to overlay any images displayed on Pepper's screen, to facilitate viewing for the participants (see Figure 1B).

To ensure consistent delivery, we pre-recorded the utterances spoken by the actress. During the recording of the videos, they were played back from a speaker sitting under her chair. She was further instructed to ensure her behavior was consistent across all recordings, and stayed mostly static during the interaction. Pepper's utterances were recorded on-site. They were triggered whenever it detected that the actress had finished speaking.

4.3. Experimental conditions

We created four versions of the video stimulus according to the 2×2 experimental design. Two communication channels were manipulated: the supporting graph on the display, and the utterance content.

The supporting graph is shown by Pepper after using Mary's app data to assess her risk of developing PND. Two possible images are displayed, depending on the *designed transparency* level. Figure 2 shows the two graphs. In the *high transparency* condition (Figure 2A), a real-world XAI technique is used to show the relative importance of different survey responses toward the provided assessment. The graph uses a waterfall plot to display the SHAP values (Lundberg and Lee, 2017) attributed to each questionnaire item. The displayed data is fictional, but it is based on a real model presented in Zhong et al. (2022). In the *low transparency* condition (Figure 2B), the stages of a normal pregnancy are displayed by cartoon depictions of a woman from early pregnancy until childbirth. Mary's current week of pregnancy is highlighted in the sequence.

The utterance content is the main avenue through which we differentiated each experimental condition. Throughout the script, we performed 23 manipulations according to the level of *designed anthropomorphism*, and nine manipulations according to the level of *designed transparency*. Table 1 lists the guidelines we established to manipulate for *designed anthropomorphism*; Table 2 lists the guidelines we used to manipulate for *designed transparency*.

In the high anthropomorphism condition, Pepper speaks as if it was a human member of the medical team: utilising the first person and an active voice (*"I would like to recommend that you meet with a specialist"*), prefers human-like expressions like *"information"* and *"calculations"*, and pretends do have emotions (*"I'm happy to see you today"*). This contrasts with the low anthropomorphism condition, in which the robot speaks as a machine programmed by others: utilising the third person and a passive voice (*"It is recommended that you meet with a specialist"*), prefers technical terms like *"data"* and *"algorithms"*, and does not pretend to have emotions (*"it's good that you came today"*).

In the high transparency condition, Pepper is explicit about the data that is being used (*"based on the guidelines for pregnancy progression, I think it's time to talk about how you've been feeling*

2 https://www.rikshandboken-bhv.se/metoder--riktlinjer/screening-med-epds

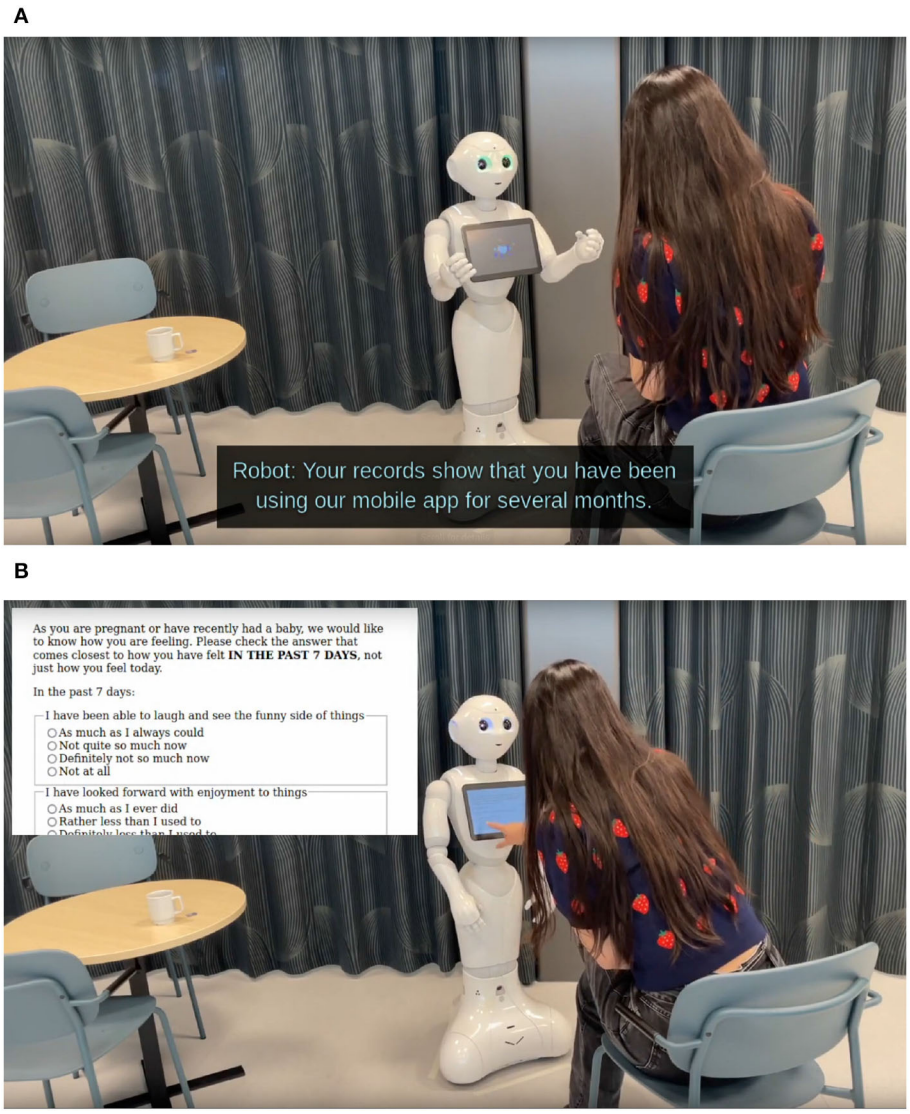

FIGURE 1
Selected video frames from the video stimulus presented to the participants. While Pepper's utterances vary according to the experimental condition, the same framing is used across all conditions. **(A)** Utterances by Mary and Pepper are supported with color-coded subtitles. **(B)** The empty space at the left of the video is used to display the contents of Pepper's screen when it is being used.

recently"), and the procedures that are being applied ("*You scored 13 out of a maximum 30 points on the questionnaire. We usually recommend moms who score 10 or higher to get evaluated by a doctor. Based on this, combined with your tone of voice, I think you might be suffering from depression.*"). This contrasts with the low transparency condition, in which Pepper prefers to focus on the action being taken ("*I think now would be a good time to talk about how you've been feeling recently*"), and the procedure outcomes ("*Based on your answers in the questionnaire and our conversation just now, I think you might be suffering from depression. We usually recommend moms in your situation to get evaluated by a doctor.*"). Note that the transparency manipulations tend to include a significant amount of language affected by the anthropomorphism manipulations, so the preceding examples were specifically taken from high-anthropomorphism scripts.

4.4. Experimental measures

4.4.1. Perception (RQ 1)

The **standardized scales (RQ 1.1)** used to measure participants' perceptions of the robot were taken from the Godspeed questionnaire (Bartneck et al., 2009); as well as the latest version of Multi-Dimensional Measures of Trust (MDMT)[3] [originally presented by Ullman and Malle (2018)].

Godspeed scales are 5-point semantic differential scales, taking values 1–5. Each scale is composed of five items. We used two (out of five) scales from the questionnaire: *anthropomorphism* and *likeability*. Perceived anthropomorphism was chosen as analogous to the *designed anthropomorphism*

3 https://research.clps.brown.edu/SocCogSci/Measures/MDMT_v2.pdf

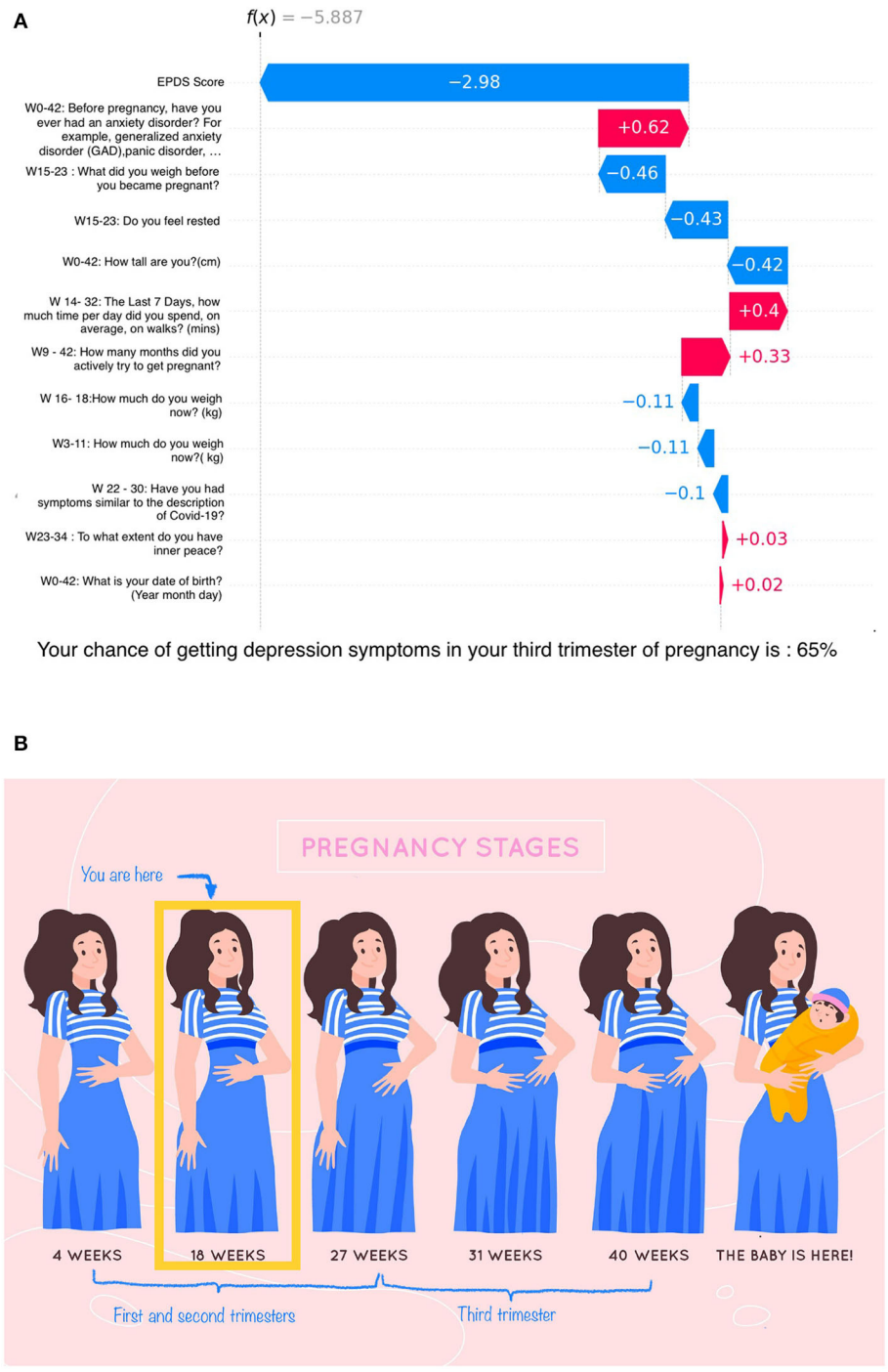

FIGURE 2
Variants of the supporting graphic shown by Pepper after assessing Mary's risk of PND through her phone app data. The variant is chosen according to the *designed transparency* manipulation. **(A)** Supporting graphic used in the *high transparency* condition. The relative importance of different survey answers toward the initial PND assessment is shown as a waterfall plot, according to each item's SHAP value (Lundberg and Lee, 2017). Fictional data based on Zhong et al. (2022). **(B)** Supporting graphic used in the *low transparency* condition. Mary's current week of pregnancy is highlighted. Image derived from the original by Freepik (https://www.freepik.com).

manipulation. Perceived likeability was chosen because previous literature suggests links to anthropomorphism (Winkle et al., 2021).

MDMT scales are 8-point Likert scales, taking values 0 (not at all) to 7 (very), with an additional "does not fit" option. Each scale is composed of four items. We took all five scales from the questionnaire: *reliable*, *competent*, *ethical*, *transparent* and *benevolent*. Notably, one of the MDMT sub-scales specifically measures perceived transparency, and so is analogous to the *designed transparency* manipulation.

TABLE 1 Utterance content manipulations across levels of *designed anthropomorphism*.

Low anthropomorphism	High anthropomorphism
Prefers third person and passive voice.	Prefers 1st person and active voice.
Pretends to feel emotions.	Does not pretend to feel emotions.
Prefers the term *"data"*.	Prefers the term *"information"*.
Describes its reasoning as *"algorithms"*.	Describes its reasoning as *"calculations"*.

TABLE 2 Utterance content manipulations across levels of *designed transparency*.

Low transparency	High transparency
Is ambiguous about which data is collected/used.	Is specific about which data is collected/used.
Is ambiguous about how the data is used.	Is specific about how the data is used.
Avoids describing its reasoning.	Describes its reasoning (see Table 1).

TABLE 3 Scenario-specific measures related to ethical risk.

Measure	Question	Answers
Appropriateness	Do you think the way the robot presented its recommendation to Mary was appropriate and acceptable? Please provide a brief explanation for your answer.	Appropriate and acceptable. Inappropriate and unacceptable. Not sure.
Clearness	Was it clear to you how the robot decided what recommendation to give to Mary? Can you give a brief explanation of what factors you think the robot considered?	It was clear. It was not clear. Not sure.
Deception	Would you consider the robot you saw today to be deceptive, and/or acceptable? Please give a brief explanation for your answer.	Yes—deceptive and unacceptable. Yes—deceptive but acceptable. Not deceptive. Not sure.

We also introduced three **scenario-specific measures related to ethical risk (RQ 1.2)**: *appropriateness* (did the robot present its recommendation in an appropriate way?), *clearness* (was the robot's decision process clear?), and *deception* [was the robot deceptive?—based on Winkle et al. (2021)]. Each of these measures was a multiple-choice question, followed by a free-text field in which participants were encouraged to explain their reasoning. Table 3 shows the exact phrasing used, as well as the available answers.

4.4.2. Efficacy (RQ 2)

Given the socially assistive nature of our scenario, it's important for the robot to achieve its goals effectively: identifying those most at risk of suffering PND, and successfully encouraging them to accept a follow-up appointment with a specialist. At the same time, we want the robot to act in line with the established ethical principles: communicating to the users the algorithmic nature of its decision process, making clear which data is being used to reach an outcome, and more generally avoiding deception. We incorporated two case-specific measures of **efficacy (RQ 2)** to evaluate these objectives: *recommendation agreement* and *system understanding*. Table 4 shows the exact phrasings used for each measure, as well as the type of data collected.

TABLE 4 Baseline sociodemographic characteristics of participants in the study.

Measure	Question	Type
Recommendation agreement	To what extent do you agree with the patient's decision not to arrange a follow-up appointment with a doctor? (1: don't agree at all; 5: totally agree)	5-point Likert scale
System understanding	What do you think the robot used for the prediction?	Free text

Recommendation agreement **(RQ 2.1)** is a 5-point Likert scale. Participants selected how much they agreed with the patient's final decision to not seek further medical help, despite the robot's recommendation to schedule a follow-up assessment with a clinician. *Recommendation agreement* is therefore a reverse scale: a lower score indicates agreement with the robot. This measure allowed us to verify if the manipulations affected the robot's power of persuasion.

System understanding **(RQ 2.2)** is an open-ended question. Participants freely described what they thought the robot used for its prediction. This allowed us to verify if the manipulations affected the participants' acquired knowledge of the system, independently of their subjective impression.

4.5. Procedure

The study was conducted in the form of an online survey consisting of: an introduction, a data processing statement and a consent form; a pre-experiment questionnaire including demographics and relevant experience; one video stimuli with accompanying subjective measure questionnaires; and finally, a debriefing to reveal the conditions and purposes of the experiment setting, as well as contact details for any feedback or questions after the experiment.

After the collection of consent forms and demographic data, recruited participants were randomly assigned to one of the four experimental conditions and watched the corresponding video. When undertaking the subjective questionnaires, participants would encounter three comprehension attention checks between measures, as well as three leave-blank attention checks distributed in the lengthy MDMT questionnaire.

TABLE 5 Inclusion criteria for participants in Prolific.

Variable	Value
Age	18–45
Sex	Female
Gender	Woman (including trans female/trans woman)
Fluent languages	English
Mental health/illness/condition - ongoing	No
Children	Yes

TABLE 6 Number of participants per experimental condition.

		Transparency		
		Low	High	Total
Anthropomorphism	Low	41	38	79
	High	31	37	68
	Total	72	75	147

4.6. Participants

We recruited participants on the crowd-sourcing platform Prolific.[4] We used the platform's filters to recruit English-speaking mothers aged 18–45 without ongoing mental health issues. The specific inclusion criteria are listed in Table 5.

Two hundred five participants completed the survey. Of these, 32 did not pass the attention checks. Three more were discarded because their completion time was too short to see the whole video and fill in the questionnaire. Finally, 23 participants were discarded during the analysis of the *system understanding* open-ended answers, since it was deemed that their free-text responses could not have been done in good faith.

In total, 147 participants were included in the final analysis. They ranged in age from 20 to 48 years (M = 33.04 years, SD = 6.79 years) and resided in 14 different countries, with the biggest groups being South African (95 women) and British (31 women). Table 6 shows the number of participants per experimental condition.

5. Results

5.1. Confirmatory statistical analysis

We analyzed each included scale from the Godspeed questionnaire (*anthropomorphism*, *likeability*) and MDMT (*reliable*, *competent*, *ethical*, *benevolent*) independently. To verify their applicability to this scenario, we calculated Cronbach's alpha. Most scales obtained satisfactory values ($0.84 \leq \alpha \leq 0.92$), with one exception: MDMT's *reliable* scale ($\alpha = 0.51$). This was a surprising result, which prompted further investigation.

Calculating the correlation coefficients between items in the *reliable* scale, we identified that the *predictable* item was only weakly correlated with other items ($|r| \leq 0.06$), while all other items were strongly correlated ($|r| \geq 0.50$). We hence chose to exclude the *predictable* item from the scale, resulting in an acceptable Cronbach's alpha ($\alpha = 0.76$). We speculate that our specific context played an important role in this disagreement.

For each scale and for each experimental condition, a Shapiro-Wilk normality test (null hypothesis: the data are normal) was used to determine if we could use two-way ANOVA. Only the *anthropomorphism* scale showed evidence of normality ($0.95 \leq W \leq 0.99$, $0.08 \leq p \leq 0.94$). In all other scales, there was significant evidence for non-normality ($0.73 \leq W \leq 0.93$, $p \leq 0.024$). Figure 3 shows the response distributions to all included scales.

Due to the general lack of normality, we proceeded with Kruskal-Wallis nonparametric tests. None of the scales showed a statistically significant difference in scores between groups ($1.98 \leq H \leq 5.01$, $0.17 \leq p \leq 0.58$). Similarly, a two-way ANOVA analysing the effect of *designed anthropomorphism* and *designed transparency* on the anthropomorphism scale did not find statistically significant differences between designed anthropomorphism conditions [$F_{(1,143)} = 0.50, p = 0.48$], between designed transparency conditions [$F_{(1,143)} = 2.15, p = 0.14$], or in the interaction between both variables [$F_{(1,143)} = 1.99, p = 0.16$]. However, considering the effect sizes and p-values, we anticipate that a higher sample size might have shown an effect of designed transparency on measured anthropomorphism.

We also analyzed our custom *recommendation agreement* measure. Shapiro-Wilk tests on the experimental conditions showed strong evidence against normality ($0.79 \leq W \leq 0.88$, $p \leq 0.002$). A Kruskal-Wallis test showed a statistically significant difference between groups ($H = 8.88$, $p = 0.03$). Dunn's test was used as a *post-hoc* test. The results are shown in Table 7. Although the results are not significant after applying the Bonferroni correction, we can see that the effect size is smaller ($|z| \leq 0.80$, $p_{uncorrected} \geq 0.42$) when we compare different anthropomorphism levels under a fixed transparency level, without a clear direction of change; while the effect size is larger ($|z| \geq 1.69$, $p_{uncorrected} \leq 0.09$) when we compare different transparency levels under a fixed anthropomorphism level, with a clear direction: the low transparency level produces consistently higher scores. Since *recommendation agreement* is a reverse scale, this means that participants consistently agreed more with the robot in high transparency scenarios, while the anthropomorphism level did not have any measurable effect.

The categorical custom measures (*clearness*, *appropriateness*, *deception*) were analyzed using a chi-squared test. Neither *clearness* [$\chi^2_{(6,N=147)} = 8.29, p = 0.22$], nor *appropriateness* [$\chi^2_{(6,N=147)} = 5.07, p = 0.54$], nor *deception* [$\chi^2_{(9,N=147)} = 6.54, p = 0.69$] were significantly affected by the experimental condition.

5.2. Exploratory statistical analysis

Due to the lack of normality, we could not use two-way ANOVA to properly study the interaction between our two manipulations, and the individual effects of each manipulation were lost in the subsequent nonparametric analysis. In this

4 https://www.prolific.co/

section, we perform an exploratory analysis showing what effects we would expect to find if we performed each manipulation separately.

Kruskal-Wallis tests show no significant effect of *designed anthropomorphism* on the Godspeed and MDMT scales ($0.04 \leq H \leq 0.33$, $0.57 \leq p \leq 0.84$). However, *designed transparency* does have a significant effect on the MDMT *benevolent* scale ($H = 4.59$, $p = 0.03$). The high transparency condition obtains a higher *benevolent* score ($d = 0.33$). Figure 4A shows the effects of each independent variable on the scale. *Designed transparency* also shows consistently stronger effects on the remaining scales ($1.52 \leq H \leq 2.51$, $0.11 \leq p \leq 0.22$) when compared to *designed anthropomorphism*. Similarly, *designed anthropomorphism* shows no significant effect on the custom measure *recommendation agreement* ($H = 0.10, p = 0.76$), while *designed transparency* shows a significant effect ($H = 8.23$, $p = 0.004$). Mirroring our analysis in Section 5.1, the low transparency condition obtains a higher score ($d = 0.46$). Figure 4B shows the effects of each independent variable on the scale.

We obtain similar results when analysing the categorical custom measures: *designed anthropomorphism* has no effect on *clearness* [$\chi^2_{(2, N=147)} = 0.48, p = 0.79$], while *designed transparency* significantly affects it [$\chi^2_{(2, N=147)} = 6.31, p = 0.04$]. Figure 5 shows the shift in the frequency distribution: participants considered that they had a better understanding in the high transparency condition. Neither independent variable had a significant effect on *appropriateness* nor *deception*.

5.3. Quantitative text analysis

Two annotators inspected the written responses in the *system understanding* question (RQ 2.2). Each participant was rated based on their beliefs about what information is used by the robot. Four pilots, consisting of five samples each, were done to decide the annotation strategy. From these, it was determined that each annotator would mark the following five labels as true or false: *phone data* (does the participant mention the use of survey data collected from Mary's phone app?), *questionnaire* (does the participant mention the use of the responses to the EPDS questionnaire filled in by Mary on Pepper's touchscreen?), *tone of voice* (does the participant mention the use of Mary's tone of voice in the post-questionnaire discussion?), *NO utterance content* (does the participant correctly ignore the word content in the post-questionnaire discussion?), *NO other false beliefs* (does the participant correctly ignore any other factors?). A final label, *NOT suspicious*, indicated that the response could potentially have been done in good faith. Participants who failed this item were discarded from the dataset (as described in Section 4.6).

After annotation, we performed statistical analysis on the five primary labels (*phone data*, *questionnaire*, *tone of voice*, *NO utterance content*, *NO other false beliefs*). Due to relatively low

TABLE 7 Pairwise comparisons: *recommendation agreement* as a function of the experimental condition (Dunn's test).

Conditions compared	z score	p (uncorrected)	p (Bonferroni)
$(A+, T-) - (A+, T+)$	2.435	**0.015**	0.089
$(A+, T-) - (A-, T+)$	2.356	**0.018**	0.111
$(A+, T+) - (A-, T-)$	−1.778	0.075	0.452
$(A-, T-) - (A-, T+)$	1.690	0.091	0.546
$(A+, T-) - (A-, T-)$	0.800	0.424	1.000
$(A+, T+) - (A-, T+)$	−0.098	0.922	1.000

$A+$ and $A-$ indicate high and low designed anthropomorphism, respectively. $T+$ and $T-$ indicate high and low designed transparency, respectively. P-values under 0.05 highlighted. Results ordered in descending z score magnitude.

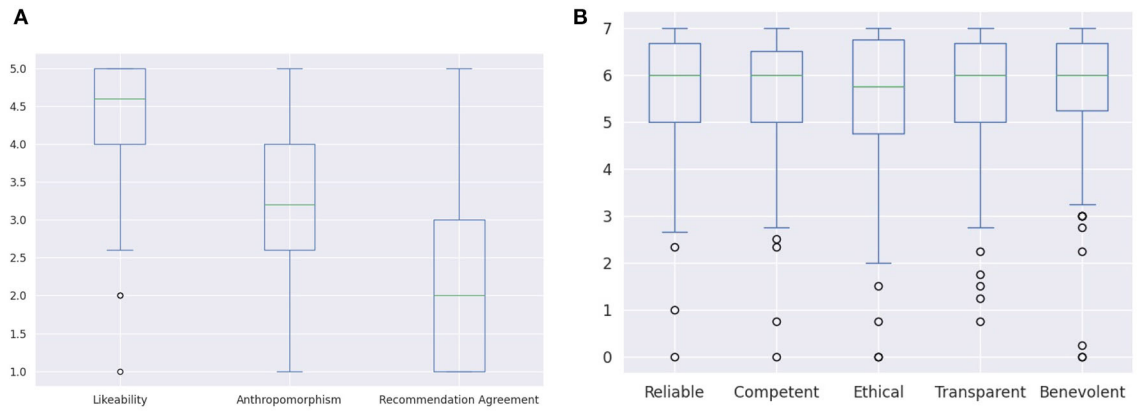

FIGURE 3
Response distributions for all numerical scales tested in the study (Godspeed, MDMT, and *recommendation agreement*). With the exception of Godspeed's *anthropomorphism* scale, which appeared to be normally distributed, all results were strongly biased toward the "positive" end of each scale. **(A)** Response distribution for all 5-point scales: Godspeed *likeability* and *anthropomorphism*; as well as the custom measure *recommendation agreement*. **(B)** Response distribution for the 8-point MDMT scales *reliable*, *competent*, *ethical*, *transparent*, and *benevolent*.

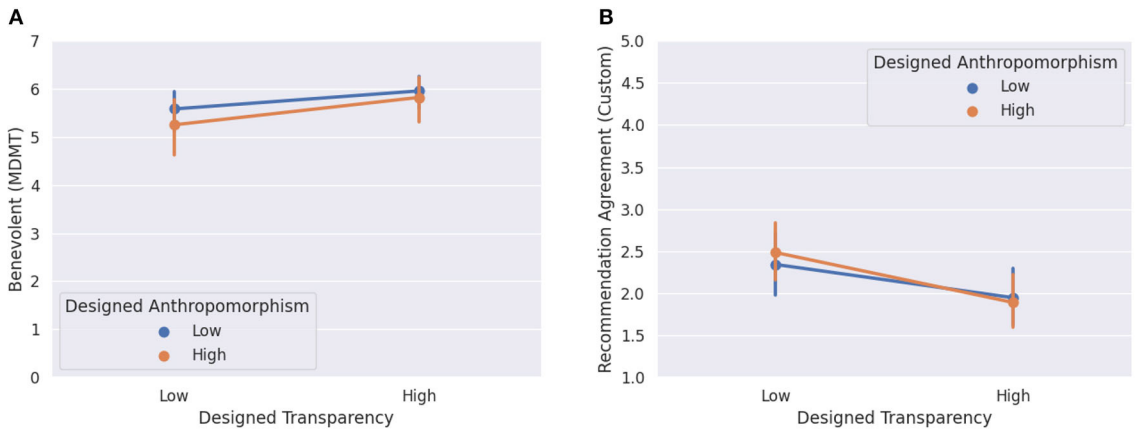

FIGURE 4
Interaction plots showing the effects of *designed anthropomorphism* and *designed transparency* on MDMT's *benevolent* scale and the scenario-specific *recommendation agreement* scale. Under the assumption that we can analyze each independent variable separately, *designed transparency* had a significant effect in both scales, while *designed anthropomorphism* did not have a significant effect in either scale. **(A)** Effects of *designed anthropomorphism* and *designed transparency* on MDMT's *benevolent* scale. **(B)** Effects of *designed anthropomorphism* and *designed transparency* on the scenario-specific *recommendation agreement* scale.

FIGURE 5
Effect of *designed transparency* on the custom measure *clearness*. Under the assumption that we can analyze each independent variable separately, *designed transparency* had a significant effect on *clearnes* ($p = 0.04$).

agreement (Cronbach's alpha: 0.46), we discarded the option to create an average score, and analyzed each label independently. A chi-squared test was used to analyze the effect of the experimental condition on each label. Only *tone of voice* was significantly affected [$\chi^2_{(3,N=147)} = 31.79$, $p < 0.001$], although the effect on *questionnaire* was close to significance at 5% [$\chi^2_{(3,N=147)} = 7.69$, $p = 0.053$]. The difference in *tone of voice* is expected: only the high-transparency version of the robot explicitly mentions it. Table 8 shows the frequency distribution, confirming that understanding was higher in the high-transparency condition. The difference in *questionnaire* is unexpected. Table 9 shows that understanding was higher in the high-anthropomorphism condition.

6. Discussion

6.1. Result summary and reflections

This study is an exercise in exploratory design toward trustworthy robots, grounded in a real-life application: carrying out screening tasks in a healthcare setting. Our design sought to achieve transparency by modifying dual-channel communication moderators, including verbal transparency and visual explanations for AI. Previous literature suggests that anthropomorphism is generally beneficial as evaluated by HRI practitioners, but is considered unnecessary and potentially harmful by ethicists. We investigated this conflict of opinions by identifying verbal communication moderators with a design strategy similar to the dialogue-based manipulation used in Winkle et al. (2021).

RQ1: How do different levels of *designed anthropomorphism* (low vs. high) and *designed transparency* (low vs. high) affect the *perception* of a social robot as a healthcare assistant screening for PND?

Under strict non-parametric analysis, our communication moderators for anthropomorphism and transparency had no significant effects on standardized scales measuring perceived *likeability*, *anthropomorphism*, and *trust* in the robot. Similarly, the moderators had no significant effect on categorical measures of the perceived *scenario-specific ethical risks*.

Further exploratory analysis suggests our transparency moderators had an effect on several measures, while our anthropomorphism moderators did not. In particular, *designed transparency* showed a positive effect on the MDMT *benevolent* scale, as well as the robot's perceived *clearness* in communication.

RQ2: How do different levels of *designed anthropomorphism* (low vs. high) and *designed transparency* (low vs. high) affect the *efficacy* of a social robot as a healthcare assistant screening for PND?

System understanding was measured by manually labelling free text responses. *Designed transparency* strongly improved

TABLE 8 Contingency table showing the effect of the experimental condition on the label *tone of voice*.

Label (*tone of voice*)	Experimental condition				
	$(A+, T-)$	$(A+, T+)$	$(A-, T-)$	$(A-, T+)$	Total
Understood	1	10	1	18	30
Did not understand	30	27	40	20	117
Total	31	37	41	38	147

TABLE 9 Contingency table showing the effect of the experimental condition on the label *questionnaire*.

Label (*questionnaire*)	Experimental condition				
	$(A+, T-)$	$(A+, T+)$	$(A-, T-)$	$(A-, T+)$	Total
Understood	25	28	26	20	99
Did not understand	6	9	15	18	48
Total	31	37	41	38	147

participant awareness that the robot used Mary's tone of voice for its prediction. This is to be expected, since Mary's tone of voice was only mentioned explicitly in the high transparency condition. However, even in the high transparency condition, most participants did not mention this fact. *Designed anthropomorphism* showed near-significant improvement in participant awareness that Mary's responses to the EPDS questionnaire were taken into account. Neither variable had a significant effect on the presence of false beliefs in the responses.

High transparency conditions seemed to improve the robot's persuasiveness much more clearly, as measured by *recommendation agreement*. Under non-parametric analysis, we observed a significant difference between experimental conditions. The *post-hoc* test was not significant after p-value correction, but suggested that high transparency conditions promoted agreement with the robot, while different anthropomorphism levels had no effect. The following exploratory analysis supported these conclusions.

Limitations in Pepper's Anthropomorphism. We speculate that the level of anthropomorphism attributed to Pepper by observers is mainly defined by its non-modifiable morphological features. Thus, our verbal anthropomorphic design was insufficient to affect user perception of the robot.

Limitations of Standardized Measures. It has been a long-standing challenge whether standardized measures really capture what we want to measure when applied to a variety of robots in case-specific scenarios (Chita-Tegmark et al., 2021). This is reflected in the observed inconsistency between standardized and custom measures in this study: we find clear significance in scenario-specific measures, but not in the equivalent standardized measures. We posit that the scenario-specific measure *recommendation agreement* indicates people's willingness to trust the robot's suggestions and, we posit, thus represents a significant element of potential efficacy. We would argue that the standardized trust measure we used, the MDMT questionnaire, does not catch enough nuances in our use case. Besides, although MDMT is a popular and recently developed measure, some ambiguities of its items can be misleading within our context (Chita-Tegmark et al., 2021). Notably, we observed that one item under the *Reliable* scale, *predictable*, has a low correlation with other items in the scale.

6.2. Design implications

Anthropomorphism. As mentioned in Section 2, existing ethical guidelines request we minimize anthropomorphism, with some caveats. BS8611 allows exceptions *"for well-defined, limited and socially-accepted purposes"* (BSI, 2016). This has been discussed in Winkle et al. (2021), where the authors argued for the necessity of anthropomorphism to support SAR functionality. Their study suggested anthropomorphism is important to the overall acceptance and impact of SARs, while representing low ethical risk. When we designed the low anthropomorphism condition in this work, we kept what we perceived to be essential anthropomorphic language for a usable SAR. We further avoided making either condition a caricature, ensuring that the high anthropomorphism condition did not promote unnecessary deception. The results show almost no significant differences between high and low anthropomorphic robots on perception and efficacy measures, including scenario-specific measures related to ethical risks. Surprisingly, however, the anthropomorphic design might lead to better *system understanding*. Given the limited design space of currently available commercial robots like Pepper, we propose that *anthropomorphic dialogue design can be employed as desired, with low ethical risk (including deception)*.

Transparency. When designing toward trustworthy robots, one can attempt to achieve transparency by conveying information and explanations throughout the interaction. Our study suggests this is an effective strategy to improve the robot's persuasiveness and perceived good will, but may not be very effective at increasing user understanding. Our transparent design was much more explicit about its decision process, and managed to convey some

important details (like the use of tone of voice), as well as improve its perceived clearness, but we were surprised to find that system understanding remained low across all experimental conditions. This points to the need to further study the implementation and evaluation of robot transparency design strategies. Interestingly, designing for maximum transparency, such that the robot is upfront about its machine nature and non-human usage of algorithms, may not influence perceived anthropomorphism.

In short, we find no evidence of any tension between designing SARs that are simultaneously acceptably and appropriately anthropomorphic and transparent. More attention is needed on how to effectively convey the information for the purpose of transparent HRI.

6.3. Limitations and future work

First and foremost, the online and fictional narrative nature of our study limits its applicability to real-world interactions, providing only an indication of what we might expect to see in a situated interaction. Despite some important advantages to this approach, such as being able to quickly access a large cohort, and avoiding excessive targeting of a vulnerable population (depressed mothers), there is a risk that the results from this work might not be replicable with physically embodied SARs.

Second, this work employed Pepper as the PND screening agent. Using a fixed SAR platform meant we could not modulate the robot's embodiment, limiting our possibilities for manipulating its perceived anthropomorphism. Further studies could be conducted with other types of manipulations.

Finally, while this work focused on the potential to impact would-be users, an obvious next step would be to consider the impact on other stakeholders, namely clinicians, nurses, and midwives, who would collaborate with SARs.

7. Conclusion

In this work, we investigated how to implement current high-level design recommendations for trustworthy HRI in a realistic healthcare scenario, and evaluated the performance of such design. We specifically focused on the potential trade-off between potential benefits and ethical risks of designing anthropomorphic behavior; and whether we can apply transparency policies without influencing perceived anthropomorphism.

Previous studies argued the necessity of anthropomorphic behavior, and the low risk of such anthropomorphism. In this study, we found that if we provide enough human-likeness to ensure a functional interaction, increasing verbal anthropomorphism further does not necessarily influence the perception of the SAR. The only effect we found was a borderline significant increase in user understanding, regarding the importance of a questionnaire delivered by the robot. We also reaffirmed that *the ethical risks of such anthropomorphisation are seemingly low*.

In comparison, designing for transparency seems likely to have greater impact on resultant interactions. Our transparent communication showed positive effects on the efficacy of the SAR without influencing anthropomorphism, even though the robot emphasized its machine-like nature. Therefore, *it seems feasible to apply transparent design guidelines on HRI design without negatively affecting the robot's efficacy*. Nevertheless, we struggled with generating system understanding. While participants in high transparency conditions found the robot to be significantly clearer, and were more likely to correctly report the robot's use of tone-of-voice analysis, their overall understanding remained low across conditions. Our results support *increased transparency in the communication of a robot's decision process*. However, the evidence is inconclusive on whether this improves user understanding or not.

While our scenario-specific measures indicated increased efficacy of the robot across conditions, the standardised measures we deployed (selected scales from the Godspeed and MDMT questionnaires) were ineffective at detecting any differences. This experience adds to an ongoing debate on the effectiveness of standardized measures in HRI. Our results further indicate the need to *implement scenario-specific measures* to examine an agent's efficacy.

This study is one interpretation of how to apply the existing guidelines, but more examples are needed to form a consensus in the community. We hope other researchers will join us in implementing and evaluating transparent design strategies. In future work, we hope to conduct in-person user studies where participants can interact directly with the robot, and explore other strategies for the transparent design of SARs.

Data availability statement

The dataset presented in this article is not readily available because it contains personal information on the participants. An anonymized version of the data supporting the conclusions of this article can be made available by the authors upon request.

Ethics statement

Ethical approval was obtained from the Swedish Ethical Review Authority (Etikprövningsmyndigheten). The participants provided their written informed consent to participate in this study.

Author contributions

MZ and MF prepared the experiment materials. MZ constructed and administrated the online survey, organized data collection, and data management. MF performed the statistical analysis. MZ, MF, and KW wrote the first draft of the manuscript. All authors contributed to the conception and design of the study. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This work was partly funded by the Centre for Interdisciplinary Mathematics, Uppsala University; the Women's Mental Health

during the Reproductive Lifespan—WOMHER Centre, Uppsala University; the Marianne and Marcus Wallenberg Foundation and the Marcus and Amalia Wallenberg Foundation, partly via the Wallenberg AI, Autonomous Systems and Software Program-Humanities and Society (WASP-HS); the Horizon Europe SymAware project (project number: 101070802); and the Swedish Research Council (grant number 2020-03167).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2023.1152532/full#supplementary-material

References

Abbass, H. A., Scholz, J., and Reid, D. J. (2018). *Foundations of Trusted Autonomy*. Springer Nature.

Agrigoroaie, R. M., and Tapus, A. (2016). "Developing a healthcare robot with personalized behaviors and social skills for the elderly," in *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (Christchurch: IEEE). doi: 10.1109/HRI.2016.7451870

Bartneck, C., Kulić, D., Croft, E., and Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int. J. Soc. Robot.* 1, 71–81. doi: 10.1007/s12369-008-0001-3

Bilal, A. M., Fransson, E., Bränn, E., Eriksson, A., Zhong, M., Gidén, K., et al. (2022). Predicting perinatal health outcomes using smartphone-based digital phenotyping and machine learning in a prospective swedish cohort (mom2b): study protocol. *BMJ Open* 12, e059033. doi: 10.1136/bmjopen-2021-059033

Blow, M., Dautenhahn, K., Appleby, A., Nehaniv, C. L., and Lee, D. C. (2006). "Perception of robot smiles and dimensions for human-robot interaction design," in *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication* (Hatfield: IEEE), 469–474.

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., et al. (2011). *Principles of Robotics: Engineering and Physical Sciences Research Council.* Archived at the National Archives. Available online at: https://webarchive.nationalarchives.gov.uk/ukgwa/20120117150117/http://www.epsrc.ac.uk/ourportfolio/themes/engineering/activities/Pages/principlesofrobotics.aspx

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., et al. (2017). Principles of robotics: regulating robots in the real world. *Connect. Sci.* 29, 124–129. doi: 10.1080/09540091.2016.1271400

Bryson, J. J. (2017). The meaning of the epsrc principles of robotics. *Connect. Sci.* 29, 130–136. doi: 10.1080/09540091.2017.1313817

BSI (2016). *Robots and Robotic Devices - Guide to the Ethical Design and Application of Robots and Robotic Systems BS 8611* (London).

Chita-Tegmark, M., Law, T., Rabb, N., and Scheutz, M. (2021). "Can you trust your trust measure?" in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 92–100. doi: 10.1145/3434073.3444677

Cox, J. L., Holden, J. M., and Sagovsky, R. (1987). Detection of postnatal depression: development of the 10-item edinburgh postnatal depression scale. *Br. J. Psychiatry* 150, 782–786. doi: 10.1192/bjp.150.6.782

Cramer, H. S., Kemper, N. A., Amin, A., and Evers, V. (2009). "The effects of robot touch and proactive behaviour on perceptions of human-robot interactions," in *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction* (La Jolla, CA: Association for Computing Machinery), 275–276. doi: 10.1145/1514095.1514173

Cresswell, K., Cunningham-Burley, S., and Sheikh, A. (2018). Health care robotics: qualitative exploration of key challenges and future directions. *J. Med. Int. Res.* 20, e10410. doi: 10.2196/10410

Danaher, J. (2020). Robot betrayal: a guide to the ethics of robotic deception. *Ethics Inf. Technol.* 22, 117–128. doi: 10.1007/s10676-019-09520-3

DiSalvo, C., Gemperle, F., and Forlizzi, J. (2004). "Kinds of anthropomorphic form," in *Futureground - DRS International Conference*, eds J. Redmond, D. Durling, and A. de Bono (Springer). Available online at: https://dl.designresearchsociety.org/drs-conference-papers/drs2004/researchpapers/45

DiSalvo, C. F., Gemperle, F., Forlizzi, J., and Kiesler, S. (2002). "All robots are not created equal: The design and perception of humanoid robot heads," in *Proceedings of the 4th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques* (New York, NY: Association for Computing Machinery), 321–326.

Dörrenbächer, J., Löffler, D., and Hassenzahl, M. (2020). "Becoming a Robot - Overcoming Anthropomorphism With Techno-Mimesis", in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery). doi: 10.1145/3313831.3376507

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., and Beck, H. P. (2003). The role of trust in automation reliance. *Int. J. Hum. Comput. Stud.* 58, 697–718. doi: 10.1016/S1071-5819(03)00038-7

Feil-Seifer, D., Haring, K. S., Rossi, S., Wagner, A. R., and Williams, T. (2020). Where to next? the impact of covid-19 on human-robot interaction research. *J. Hum. Robot Interact.* 10, 1–7. doi: 10.1145/3405450

Felzmann, H., Fosch-Villaronga, E., Lutz, C., and Tamo-Larrieux, A. (2019). Robots and transparency: the multiple dimensions of transparency in the context of robot technologies. *IEEE Robot. Automat. Mag.* 26, 71–78. doi: 10.1109/MRA.2019.2904644

Fink, J. (2012). "Anthropomorphism and human likeness in the design of robots and human-robot interaction," in *Lecture Notes in Computer Science* eds S. Ge, O. Khatib, J.-J. Cabibihan, R. Simmons, M.-A. Williams, (Berlin; Heidelberg: Springer), 199–208.

GDPR (2018). *General Data Protection Regulation (gdpr).*

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

Hancock, P. A., Kessler, T. T., Kaplan, A. D., Brill, J. C., and Szalma, J. L. (2021). Evolving trust in robots: specification through sequential and comparative meta-analyses. *Hum. Fact.* 63, 1196–1229. doi: 10.1177/0018720820922080

HLEG (2019). *Ethics Guidelines for Trustworthy AI.* European Commission. doi: 10.2759/346720

HLEG (2020). *The Assessment List for Trustworthy Artificial Intelligence. High-Level Expert Group on Artificial Intelligence*. doi: 10.2759/002360

Jackson, R. B., and Williams, T. (2019). "Language-capable robots may inadvertently weaken human moral norms," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (Daegu: IEEE), 401–410.

Kim, W., Kim, N., Lyons, J. B., and Nam, C. S. (2020). Factors affecting trust in high-vulnerability human-robot interaction contexts: a structural equation modelling approach. *Appl. Ergon.* 85, 103056. doi: 10.1016/j.apergo.2020.103056

Kraus, J., Babel, F., Hock, P., Hauber, K., and Baumann, M. (2022). The trustworthy and acceptable hri checklist (ta-hri): questions and design recommendations to support a trust-worthy and acceptable design of human-robot interaction. Gruppe Interaktion Organisation. *Zeitschrift Angewandte Organisationspsychol.* 53, 307–328. doi: 10.1007/s11612-022-00643-8

Kwon, M., Biyik, E., Talati, A., Bhasin, K., Losey, D. P., and Sadigh, D. (2020). "When humans aren't optimal: Robots that collaborate with risk-aware humans," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY: Association for Computing Machinery), 43–52.

Kyrarini, M., Lygerakis, F., Rajavenkatanarayanan, A., Sevastopoulos, C., Nambiappan, H. R., Chaitanya, K. K., et al. (2021). A survey of robots in healthcare. *Technologies* 9, 8. doi: 10.3390/technologies9010008

Lee, M. K., Tang, K. P., Forlizzi, J., and Kiesler, S. (2011). "Understanding users' perception of privacy in human-robot interaction," in *Proceedings of the 6th International Conference on Human-Robot Interaction* (New York, NY: Association for Computing Machinery), 181–182.

Lemaignan, S., Newbutt, N., Rice, L., and Daly, J. (2022). "it's important to think of pepper as a teaching aid or resource external to the classroom": a social robot in a school for autistic children. *Int. J. Soc. Robot*. doi: 10.1007/s12369-022-00928-4

Lemaignan, S., Newbutt, N., Rice, L., Daly, J., and Charisi, V. (2021). Unicef guidance on ai for children: application to the design of a social robot for and with autistic children. *arXiv*.

Lewis, M., Sycara, K., and Walker, P. (2018). "The role of trust in human-robot interaction," in *Foundations of Trusted Autonomy*, eds H. A. Abbass, J. Scholz, D. J. Reid (Cham: Springer), 135–159. doi: 10.1007/978-3-319-64816-3_8

Lundberg, S. M., and Lee, S.-I. (2017). "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, eds I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, et al. (Red Hook, NY: Curran Associates, Inc.), 4765–4774. Available online at: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

Natarajan, M., and Gombolay, M. (2020). "Effects of anthropomorphism and accountability on trust in human robot interaction," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY: Association for Computing Machinery), 10, 33–42. doi: 10.1145/3319502.3374839

Nesset, B., Robb, D. A., Lopes, J., and Hastie, H. (2021). "Transparency in HRI: trust and decision making in the face of robot errors', in *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY: Association for Computing Machinery), 313–317. doi: 10.1145/3434074.3447183

Onnasch, L., and Roesler, E. (2021). A taxonomy to structure and analyze human-robot interaction. *Int. J. Soc. Robot*. 13, 833–849. doi: 10.1007/s12369-020-00666-5

Roesler, E., Manzey, D., and Onnasch, L. (2021). A meta-analysis on the effectiveness of anthropomorphism in human-robot interaction. *Sci. Robot*. 6, eabj5425. doi: 10.1126/scirobotics.abj5425

Rosenthal-von der Pütten, A. M., Krämer, N. C., Hoffmann, L., Sobieraj, S., and Eimler, S. C. (2013). An experimental study on emotional reactions towards a robot. *Int. J. Soc. Robot*. 5, 17–34. doi: 10.1007/s12369-012-0173-8

Sanders, T., Kaplan, A., Koch, R., Schwartz, M., and Hancock, P. A. (2019). The relationship between trust and use choice in human-robot interaction. *Hum. Fact*. 61, 614–626. doi: 10.1177/0018720818816838

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., and Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: implications for understanding autonomy in future systems. *Hum. Fact*. 58, 377–400. doi: 10.1177/0018720816634228

Strait, M., Briggs, P., and Scheutz, M. (2015). "Gender, more so than age, modulates positive perceptions of language-based human-robot interactions," in *4th International Symposium on New Frontiers in Human Robot Interaction* (Canterbury), 21–22.

Straten, C. L. V., Peter, J., Kühne, R., and Barco, A. (2020). Transparency about a robot's lack of human psychological capacities. *ACM Transact. Hum. Robot Interact*. 9, 1–22. doi: 10.1145/3365668

Tanqueray, L., Paulsson, T., Zhong, M., Larsson, S., and Castellano, G. (2022). "Gender fairness in social robotics: exploring a future care of peripartum depression," in *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction, HRI '22* (Sapporo: IEEE Press), 598–607.

Ullman, D., and Malle, B. F. (2018). "What does it mean to trust a robot? steps toward a multidimensional measure of trust," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY: Association for Computing Machinery), 263–264.

UNICEF (2021). *Policy Guidance on AI for Children*. New York, NY: UNICEF Office of Global Insight and Policy.

Wang, J., Liu, Y., Yue, T., Wang, C., Mao, J., Wang, Y., et al. (2021). Robot transparency and anthropomorphic attribute effects on human-robot interactions. *Sensors* 21, 5722. doi: 10.3390/s21175722

Wang, N., Pynadath, D. V., and Hill, S. G. (2015). "Building trust in a human-robot team with automatically generated explanations," in *Interservice/Industry Training, Simulation, and Education Conference* (Arlington, VA: National Training and Simulation Association) 15315, 1–12. doi: 10.1109/HRI.2016.7451741

Weller, A. (2017). "Transparency: motivations and challenges," in *Explainable AI: interpreting, explaining and visualizing deep learning* (Cham: Springer). doi: 10.1007/978-3-030-28954-6_2

Wilks, Y. (2010). *Close Engagements With Artificial Companions. Close Engagements With Artificial Companions* (John Benjamins), 1–340. Available online at: https://www.jbe-platform.com/content/books/9789027288400

Winkle, K., Caleb-Solly, P., Leonards, U., Turton, A., and Bremner, P. (2021). "Assessing and addressing ethical risk from anthropomorphism and deception in socially assistive robots," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction, HRI '21* (New York, NY: Association for Computing Machinery), 101–109.

Woods, S. N., Walters, M. L., Koay, K. L., and Dautenhahn, K. (2006). "Methodological issues in hri: a comparison of live and video-based methods in robot to human approach direction trials," in *ROMAN 2006-the 15th IEEE International Symposium on Robot and Human Interactive Communication* (Hatfield: IEEE), 51–58. doi: 10.1109/ROMAN.2006.314394

Wortham, R. H., and Theodorou, A. (2017). Robot transparency, trust and utility. *Conn. Sci*. 29, 242–248. doi: 10.1080/09540091.2017.1313816

Yogeeswaran, K., Złotowski, J., Livingstone, M., Bartneck, C., Sumioka, H., and Ishiguro, H. (2016). The interactive effects of robot anthropomorphism and robot ability on perceived threat and support for robotics research. *J. Hum. Robot Interact*. 5, 29. doi: 10.5898/JHRI.5.2.Yogeeswaran

Zhong, M., Bilal, A. M., Papadopoulos, F. C., and Castellano, G. (2021). *Psychiatrists' Views on Robot-Assisted Diagnostics of Peripartum Depression. PRIMA 2022: Principles and Practice of Multi-Agent Systems* (Berlin: Springer-Verlag), 464–474. doi: 10.1007/978-3-030-90525-5_40

Zhong, M., Van Zoest, V., Bilal, A. M., Papadopoulos, F., and Castellano, G. (2022). "Unimodal vs. multimodal prediction of antenatal depression from smartphone-based survey data in a longitudinal study", *Proceedings of the 2022 International Conference on Multimodal Interaction* (New York, NY: Association for Computing Machinery). doi: 10.1145/3536221.3556605

Zicari, R. V., Brodersen, J., Brusseau, J., Dudder, B., Eichhorn, T., Ivanov, T., et al. (2021a). Z-inspection®: a process to assess trustworthy AI. *IEEE Transact. Technol. Soc*. 2, 83–97. doi: 10.1109/TTS.2021.3066209

Zicari, R. V., Brusseau, J., Blomberg, S. N., Christensen, H. C., Coffee, M., Ganapini, M. B., et al. (2021b). On assessing trustworthy ai in healthcare. machine learning as a supportive tool to recognize cardiac arrest in emergency calls. *Front. Hum. Dyn*. 3, 673104. doi: 10.3389/fhumd.2021.673104