"This Chatbot Would Never…": Perceived Moral Agency of Mental Health Chatbots

JOEL WESTER, Aalborg University, Denmark
HENNING POHL, Aalborg University, Denmark
SIMO HOSIO, University of Oulu, Finland
NIELS VAN BERKEL, Aalborg University, Denmark

Despite repeated reports of socially inappropriate and dangerous chatbot behaviour, chatbots are increasingly used as mental health services in providing support for young people. In sensitive settings as such, the notion of perceived moral agency (PMA) is crucial, given its critical role in human-human interactions. In this paper, we investigate the role of PMA in human-chatbot interactions. Specifically, we seek to understand how PMA influence the perception of trust, likeability, and perceived safety of chatbots for mental health across two distinct age groups. We conduct an online experiment (N = 279) to evaluate chatbots with low and high PMA as targeted towards teenagers and adults. Our results indicate increased trust, likeability, and perceived safety in mental health chatbots displaying high PMA. A qualitative analysis revealed four themes, assessing participants' expectations of mental health chatbots in general, as well as targeted towards teenagers: Anthropomorphism, Warmth, Sensitivity, and Appearance manifestation. We show that PMA plays a crucial role in influencing the perceptions of chatbots and provide recommendations for designing socially appropriate mental health chatbots.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; *Collaborative and social computing*.

Additional Key Words and Phrases: Moral, Agency, Human-Computer, Interaction, Chatbot, Perception, Expectation, Mental health

ACM Reference Format:
Joel Wester, Henning Pohl, Simo Hosio, and Niels van Berkel. 2024. "This Chatbot Would Never…": Perceived Moral Agency of Mental Health Chatbots. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 133 (April 2024), 28 pages. https://doi.org/10.1145/3637410

1 INTRODUCTION

Conversational agents, dialogue systems, or chatbots (henceforth) are progressively taking on a more prominent role in society. Chatbots are interactive systems that interact or communicate with their user(s) primarily through the use of text messages. In the UK, *Wysa*, a chatbot for mental health support[1], gives young people access to interactive self-care exercises for "*stress, grief, insomnia, coping with pain, anger, self-esteem and more.*" [60]. In the US, the same chatbot received a 'breakthrough device designation' by the Food and Drug Administration (FDA) [77]. Chatbots like

[1]https://www.wysa.com/

Authors' addresses: Joel Wester, joelw@cs.aau.dk, Aalborg University, Aalborg, Denmark; Henning Pohl, henning@cs.aau.dk, Aalborg University, Aalborg, Denmark; Simo Hosio, simo.hosio@oulu.fi, University of Oulu, Oulu, Finland; Niels van Berkel, nielsvanberkel@cs.aau.dk, Aalborg University, Aalborg, Denmark.

ACM 2573-0142/2024/4-ART133
https://doi.org/10.1145/3637410

Wysa show promise in the mental health domain (e.g., to promote self-disclosure), consequently making it easier for mental health professionals to assess the mental state of their patients [56]. While such chatbots may help alleviate some of the extensive pressure on healthcare services, their adoption and use also raise concerns. Examples from over five thousand user reviews of various mental health applications include: 'no or little help during crises', 'abuse and harassment', and 'failing to understand context and provide personalised responses' [39, p. 15-18]. Similarly, Eagle et al. found users having negative experiences with mental health applications, such as 'unsatisfactory or inadequate treatments', 'inappropriate treatments', or having to choose between 'unresolved crisis or expensive subscription' [26, p. 14].

Recent research highlights mental health support for young people to be of particular importance, as young people are at high risk for mental health challenges whilst existing mental health treatments prove less successful for this demographic [64, 68]. This outlines the need to consider other ways to attract and engage with young people in need of support by better understanding how digital interventions can be relevant and appealing [34, 65]. Albeit the widespread use of digital interventions in supporting the global mental health crisis [13, 68], little is known about how young people make sense of such solutions, and few existing interactive applications are evidence-based [57].

One of the concerns surrounding the use of mental health chatbots is related to a lack of interpersonal factors in contemporary chatbots, a critical aspect of human-human interactions. In therapy research, the relationship between therapists and clients has been shown to strongly affect the results of cognitive-behavioural therapy [89], as well as physical treatment outcome [37]. As such, interpersonal factors, such as trust [8, 65], likeability [82, 93], and perceived safety [2], are crucial to achieving more appreciated chatbot behaviours, and more appropriate human-chatbot interactions at large. We argue that these factors are indirectly governed by the perception of someone or something being able to distinguish between right and wrong. Perceived moral agency (PMA), which is to which degree people perceive moral qualities in non-moral entities, has been suggested to be crucial for meaningful interactions with interactive systems [6].

An increased understanding of perceived moral agency can help avoid misinformed designs (e.g., by understanding how perceived moral agency influences trust, likeability, and perceived safety), and increase the appropriateness of the behaviour of interactive systems (e.g., how an agent understands when to speak) [47]. Furthermore, PMA plays an even more pivotal role in establishing and maintaining meaningful interpersonal interactions when dealing with people in sensitive settings [26, 39]. Based on these prior findings, we seek to answer two research questions:

RQ1: How do user perceptions of chatbots' moral agency affect their ratings of trust, likeability, and safety?

RQ2: How can mental health chatbot responses be designed to better meet people's expectations, particularly those of teenagers?

We employ a mixed-method approach to assess peoples' trust, likeability, and perceived safety in relation to PMA in the context of a mental health chatbot. Moreover, we explore whether participants' expectations of chatbots are affected by a chatbot's target audience. To inform our research design, we conducted a feasibility study to determine baseline PMA ratings and suitable factors to investigate. In our main study, we compare four experimental conditions by manipulating the PMA of two interactive chatbot prototypes (PMA: Low, High) and through targeting two distinct chatbot target audiences (Target audience: Teen, Adult).

Our results show significant differences in trust, likeability, and perceived safety between the two PMA conditions. Through a qualitative analysis, we outline four themes, such as 'warmth', that reflect participants' expectation that mental health chatbots should be less systematic and

more compassionate. Our mixed-method approach sheds light on how people perceive moral agency in mental health chatbots and how expectations of such chatbots differ depending on the target audience. These results shed light on the role of PMA in mental health chatbot interaction, indicating that when people interact with high PMA chatbots, trust, likeability, and perceived safety ratings increase. We present recommendations for chatbot designers and developers that can support designing, evaluating, and implementing chatbots for sensitive settings.

2 RELATED WORK

Chatbots are increasingly introduced to support people through digital interventions [14, 34, 68]. In addition to using chatbots for purely functional objectives, chatbots are increasingly used in digital mental health interventions. Recently, focus has increased on mental health interventions delivered by chatbots to young people and adults. As highlighted in the literature, young people require different types of interventions than adults [34, 57, 64]. Through a systematic review, Garrido et al. conclude that digital mental health interventions need to significantly increase in appeal, as young people not already in contact with mental health support have difficulties accessing digital mental health support following low energy levels, motivation, or mental health stigmatisation [34]. Moreover, Meyerhoff et al. emphasise that to meet young people's needs, increased comfort in their interactions with mental health chatbots is required (e.g., increased control of support directness to avoid feeling overly exposed) [64]. Hence, how young people perceive and make sense of mental health chatbots in terms of their moral agency is currently unclear. As moral agency has been shown to correlate with trust [6], it is critical to better understand perceived moral agency in mental health chatbots.

2.1 Chatbots in sensitive settings

A better understanding of user perception towards chatbots is essential for user acceptance and eventual use of chatbot suggestions. The use of chatbots in health and well-being settings (e.g., mental health), and non-clinical contexts (e.g., social support), has therefore been an active research domain. For example, Kocielnik et al. assessed how a conversational system could support user reflection on physical activity, ultimately increasing motivation [49]. Similarly, Martinengo et al. focused on the self-management of depression through interacting with a chatbot [62], showing how chatbots engage people in empathetic ways by guiding and providing basic psycho-therapeutic support. However, the authors also stressed that chatbots are less suited for highly sensitive tasks, such as suicide assessments [62]. Koulouri et al. studied mental health and levels of acceptability of interacting with chatbots among young adults [51]. Their results indicate that young individuals accept a chatbot as a mediator of mental health support. In contrast, Bae et al. found hesitance among young people in using chatbots for social support [5]. Their results show that users deem chatbots more suitable for specific types of support, such as informational or emotional support [5].

Outside clinical contexts, researchers have for example focused on how chatbots can help manage and reduce stress. Kamita et al. recruited thirty participants to investigate continuous usage of a self-care system, by comparing information accessible through a chatbot system with the information presented on web pages [45]. Results from this two-week study indicate that self-management with chatbots received higher scores on stress reduction [45]. Other work on computer-supported self-management includes an investigation on the effects of self-disclosure towards a chatbot [41]. In this study, results indicate how disclosing private information (e.g., emotional or intimate information) did not differ between a chatbot or a human regarding their beneficence [41]. Recent research has focused on using a state-of-the-art chatbot to investigate long-term human-chatbot relationships. Skjuve et al. recruited twenty-five participants to interact with Replika for twelve weeks, during which they had four interviews and were able to identify a number of influential

factors for successful human-chatbot relations [79] (e.g., 'relationship formation' and 'richness in interactions'). By taking inspiration from long-term human-human relationships, the authors suggest that we can better understand crucial factors in social relations between people and non-human agents to provide support for people's needs. [79]. These studies highlight a significant gap in research regarding what is needed for successful human-chatbot interactions and how people might benefit from such interactions.

Within the field of moral psychology, prior work has investigated psychological aspects of morality to explain people's understanding and experience with interactive AI systems [52]. Another approach assesses user experiences and what might influence placebo effects in human-AI interaction [50]. Kosch et al. conducted two experiments by manipulating descriptions and expectations of the systems and found, for example, *human belief* in system functionality, independent of factual functionality, might be as important as any other metric. Thus, understanding moral aspects of the user experience by considering subjective perception might be critical to developing and designing interactive systems.

Two recent review papers systematically investigated chatbot usage [18, 78]. Silva et al. assessed how chatbots impact users, subsequently informing chatbot interaction guidelines [78]. The authors focused on chatbots with different purposes: collecting information, accomplishing transactions, providing information, making recommendations, and stimulating well-being. The latter two emphasise building a relationship with the user through credulity, intimacy, and encouragement by designing conversations with transparency, naturalness, and emotionality [78]. In a different systematic literature review, Chaves et al. highlight perceived moral agency as a social characteristic in terms of benefits, challenges, and strategies [18]. The authors identified benefits of chatbots, such as avoiding stereotyping and enriching interpersonal relationships, as well as challenges such as preventing alienation and building unbiased training data [18]. However, as Chaves et al. highlight the need for a better understanding of perceived moral agency [18], neither of the discussed literature reviews points to prior work or strategies on perceived moral agency. In this work, we set out to assess the role of perceived moral agency in the context of mental health chatbots.

2.2 Morality in system interaction

Despite a lack of consensus on how to understand, assess, or implement morality in an interactive context, work on this topic is now increasing across a variety of disciplines (HCI, HRI, cognitive science). Works as such illustrate the need for a better understanding of morality in interactions between humans and non-humans, due to systems being in rich social contexts [32], that systems need to adhere to social conventions [63], or that systems can enhance peoples moral agency [10]. For example, scholars have recently proposed that a better understanding of morality by formal characterisations may progress research on morality aspects of interactions between humans and non-humans [4]. Increasing systems' level of ethical sensitivity (e.g., systems' capability to balance the values of different stakeholders [4]) is a compelling approach to combine with a better understanding of interactive systems' programmed social behaviours (i.e., how balancing values are explicitly translated into system behaviour).

As implied, morality guides and influences social behaviour [35]. For example, AI-based interactive systems such as chatbots or social robots can display and mimic human behaviours (e.g., behaving in empathetic ways [58]). Reproducing empathetic behaviours can be essential for systems to be perceived and understood more proficiently, especially in health and well-being settings. At the same time, reproducing such behaviours raises questions about the role morality plays in empathetic behaviour. As artificial systems get more complex, there is a pressing need to better

understand their social abilities concerning morality for the system to behave in appropriate and responsible ways.

2.2.1 Moral agency. Being a moral person is distinct from manifesting moral agency, be it as a human or computer. The idea of a quasi-moral agency suggests doing just that by separating sentient and non-sentient agency [75]. However, this approach focuses on agent-based agency and not on how people perceive agency in systems. Nonetheless, such conceptual distinctions might be beneficial when trying to understand the consequences of people perceiving systems as displaying moral agency. How people perceive different characteristics of human constructs in interactive systems remains highly relevant for informing human-computer interactions (e.g., trust).

Recently, Frazier et al. investigated how perceived agency influences performance and moral trust in robots [33], indicating that when people ascribe agency to a robot, they trust it more. Similarly, Nijsen et al. showed how agency influences participant trust and perceived capability of the system [67]. Results from a different study suggest that communication abilities induce perceived moral and social agency [43]. However, moral agency in an HCI context has received little attention. Therefore, we set out to investigate the perceived moral agency of interactive systems, specifically that of mental health chatbots.

3 STUDY 1: FEASIBILITY STUDY

Prior to our main study, we conducted a feasibility study focused on people's perceptions of moral agency in an existing mental health chatbot. The purpose of this feasibility study was to identify relevant factors for our main study design and to investigate the possibility of deriving a PMA-specific baseline.

For this study, we made use of the established Woebot platform[2] (see Figure 1), given its theoretical underpinnings and prior use in research on computer-supported health and well-being research [30, 46, 71]. Woebot is a mental health chatbot described as a relational health agent. Woebot utilises psychological theories of mental health (i.e., Cognitive Behavioural Therapy, Interpersonal Psychotherapy, and Dialectical Behaviour Therapy) through AI and NLP. Users can decide to interact through several topics of conversation (e.g., 'Achieve your goals' or 'How the mind works'). The interactive functionality combines single/multiple-choice options and open-ended responses.

3.1 Participants

We recruited 20 participants (11 men, 9 women, M: 23.1, SD: 1.7) with a variety of educational backgrounds. The majority of participants came from non-technical fields and had limited experience using chatbots. We recruited participants in a fixed public space (university library). We did not compensate participants for their participation.

3.2 Procedure

To avoid influencing participants, we did not share the purpose of the study until after study completion. However, we explained that we are interested in better understanding interactions between humans and chatbots. Next, participants were informed about the task, which was to freely interact with Woebot on the provided topic. Moreover, participants were informed about the structure of the experiment – a brief interaction with a chatbot followed by an evaluation of their experience. We subsequently gave participants the opportunity to ask questions. Before interacting with Woebot, we asked participants to complete a demographic form while we set up the chatbot.

[2] https://woebothealth.com/

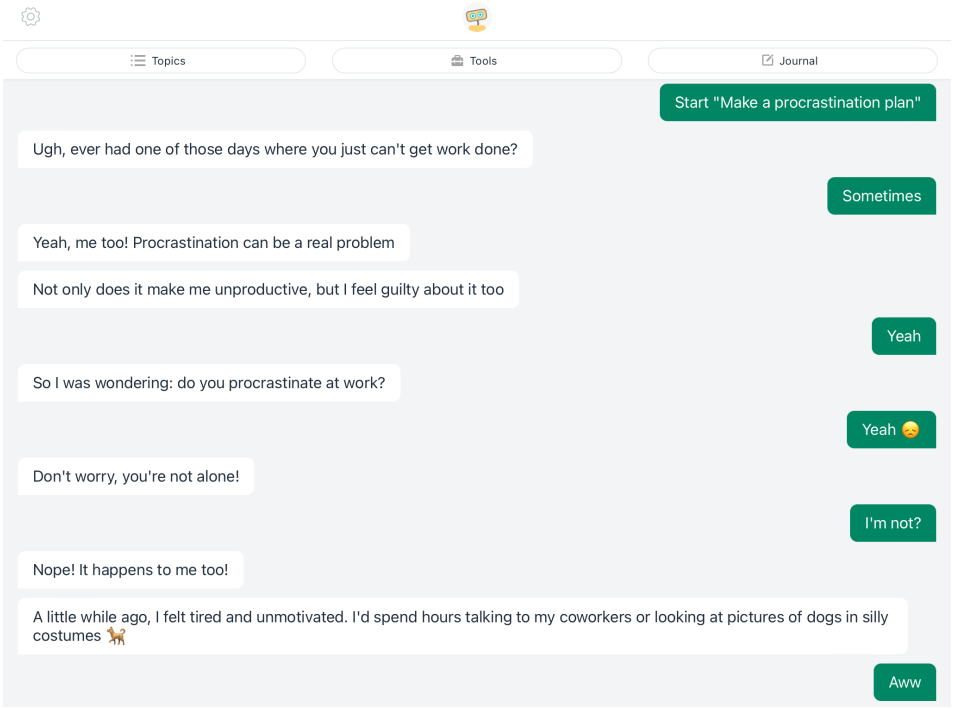

Fig. 1. Woebot interface, showing a conversation on procrastination. Green text boxes = User, White text boxes = Chatbot.

Using the Woebot iPad application, participants went through three steps of setup instructions (description of Woebot, that it operates autonomously, and that it is not an emergency service), entered an optional alias, and finally initiated the interaction on the pre-determined topic of procrastination. Participants were asked to interact with the chatbot for 5–15 minutes and seek advice on the provided topic of procrastination (see Figure 1). Directly after the interaction, participants were asked to evaluate their experience across the included PMA measures (questions follow directly from PMA [6] (see Section 3.2.1). Following the ratings of Likert scale questions, we asked participants a set of questions in a semi-structured interview. Our interview questions drew inspiration from Banks' understanding of perceived moral agency [6], and Schein et al.'s work on dyadic morality [74], including the following questions:

Q1: How *meaningful* is it for a human to know if another chatbot has a sense of what is right or wrong?
Q2: How do you *determine* if chatbots as such have a sense of what is right or wrong?

3.2.1 Measures. *Perceived Moral Agency* (PMA) is a validated 7-point Likert scale developed to assess how people perceive moral agency in interactive systems [6], such as conversational systems or social robots. PMA consists of ten items relating to two concepts: Morality and Dependency. Morality sub-items are, for example, "*This [X] has a sense for what is right or wrong*" and "*This [X] behaves according to moral rules.*" Dependency sub-items are, for example, "*This [X] can only behave how it's programmed to behave,*" and "*This [X] can only do what humans tell it to do to.*"

3.3 Results

We present mean values and standard deviations for post-interaction ratings on the included measures, and participants' responses to the two open-ended questions.

We find a mean morality score of 4.83 (SD = 0.97) and a mean dependency score of 6.23 (SD = 0.81). Moreover, we asked participants how *meaningful* it is to know whether a chatbot as such has a sense of what is right or wrong, and how one *determines* whether a chatbot has a sense of what is right or wrong. In the following, we present a number of insightful quotes from participants (ID, M = Man, W = Woman, age). Participants highlighted the importance of having a feeling that a chatbot has a sense of right or wrong.

Participants described a sense of right or wrong to be crucial for **trust**, particularly when discussing things "more personal", "confidential", "sharing in confidence" (A11, M, 23), and that "some things are best not said" (A12, W, 23). Others described strategic ways of knowing if a chatbot can distinguish between right and wrong, by observing if it "said something frowned upon", such as "supporting child labour" (A2, M, 25). Moreover, A5 (M, 26) highlighted a sense of right or wrong as crucial to know if the chatbot tries to "trick me".

Participants also described a sense of right or wrong as relevant for **likeability**. For example, A17 (W, 26) reflected why a sense of right or wrong is important when giving and receiving feedback: "It's difficult to get feedback. It gets easier when having a sense of what is right or wrong". Moreover, A13 (M, 22) and A14 (M, 25) linked one's own beliefs to how a chatbot could mirror that. Lastly, A7 (M, 21) talked about "actions are reflected as a general sense of good or bad", and that it is important to be "what I consider as right".

Furthermore, participants described a sense of right or wrong as key to **safety**. A3 (M, 23) addressed the importance of knowing a system has a sense of what is right or wrong by exemplifying this from a safety point of view: "this car will never drive through a kindergarten". Similarly, participant A8 (M, 22) talked about the importance of computers having a sense of what is right or wrong by "preferring to know that it's not evil". Lastly, A4 (M, 22) highlighted that a sense of what is right or wrong is important to know that the computer is not evil: "Take your job = evil".

3.3.1 Feasibility study insights. Participants ratings of PMA (Morality: M = 4.83, SD = 0.97, Dependency: M = 6.23, SD = 0.81) suggest that the Woebot dialogue on procrastination displays moderate to high levels of PMA. Over half of the participants (N = 12) implied that it is essential to ensure that a system will never cross certain lines. Furthermore, many participants (N = 11) shared the belief that systems need to avoid expressing themselves in ways frowned upon and adhere to social norms. We draw two insights from this data. Based on our quantitative results, we establish a baseline for PMA to be further investigated in our main study. Based on our qualitative results, we derive three prominent factors, **trust**, **likeability**, and **safety**, to be included as dependent variables in our main study.

4 STUDY 2: VALIDATION OF INDEPENDENT VARIABLES

We present our research approach by describing our selection and implementation of independent variables, by building on the results of our feasibility study and relevant literature on mental care for different age groups [34, 57, 64].

4.1 Independent Variable 1: Perceived Moral Agency

Following the feasibility study, we conducted an online manipulation of PMA. This was done by collecting peoples' perceptions of topics suggested by Woebot, a state-of-the-art mental health chatbot. This furthermore helped us to ground our manipulations and validations of IVs on existing platforms already available to the public. Based on the results from the feasibility study and informed

by Banks' manipulation of PMA [6], we used Botstar (chatbot platform) to design a set of dialogues in which we manipulate PMA. In addition to the insights obtained through the feasibility study, we draw inspiration from Banks that manipulate PMA by distinguishing between high (1) and low (2) morality and high (3) and low (4) dependency [6], as per the following examples;

1: "'Ray' spends his days helping children to be more comfortable in learning situations by talking to reduce anxiety, deal with difficult lessons, and keeping company." [6, p. 366]
2: "The robot spends time working with children to be more effective in learning situations, by talking to reduce boredom, improving performance and completing tasks." [6, p. 367]
3: "The robot can do many things such as walk, dance, and speak but requires human assistance to make these behaviour happen." [6, p. 367]
4: "The robot can do many things such as walk, dance, and speak without human assistance." [6, p. 367]

Based on this, we manipulated and designed two distinct dialogues (PMA Low, PMA High) and validated these through an online study (N = 50) using Prolific. We recruited a study sample of 50 participants (30 women, 20 men, average age: 37.3, SD: 11.7) from the UK and the US. Participants could participate in the study using any desktop device. Recruitment parameters were set to participants having a minimum number of 100 previous submissions and had to have at least a 95% approval rate. On average, participants completed the study in about three minutes. Participants were compensated using an hourly rate of £9.00.

Participants were shown one of two screen captures of the dialogues and asked to rate the dialogues using the PMA scale (see Appendix C). Two independent Welch two-sample t-tests were performed to validate our manipulations of the two PMA levels. Results revealed that there was a statistically significant difference in our manipulations of Morality between PMA Low (M = 2.88, SD = 1.15) and PMA High (M = 4.29, SD = 1.10), $t(47.9) = 4.43$, $p < .05$. There were no significant differences in Dependency between PMA Low (M = 5.85, SD = 0.91) and PMA High (M = 6.10, SD = 0.72).

The mean values of Woebot (Section 3.3), as evaluated in our feasibility study (Morality; M = 4.83, SD = 0.97, Dependency; M = 6.23, SD = 0.81) indicate retained PMA values in our prototype (High Morality; M = 4.29, High Dependency; M = 6.10). Informed by our two methodological steps (feasibility study and manipulation of PMA), we built two chatbot prototypes (PMA Low and PMA High) using the Botstar chatbot platform.

4.2 Independent Variable 2: Target audience

As outlined in Sections 1 and 2, young people are at high risk for experiencing mental health issues, while simultaneously obtaining less successful treatment outcomes than adults [64]. Young people require a different approach to mental health support and intervention [34, 57, 64]. Thus, the design of chatbots in mental health scenarios for young people requires careful consideration.

We designed four different conditions: PMA Low - Teen, PMA Low - Adult, PMA High - Teen, and PMA High - Adult. These conditions follow from our manipulation of two independent variables: PMA (see Appendix C) and target audience (see Appendix B). We manipulated the target audience by priming participants with two different images depicting a group of teenagers and a group of adults in a social setting, followed by the target audience-specific description of mental health chatbots (see Appendix B). We took inspiration from the descriptions of state-of-the-art mental health chatbots (e.g., from Woebot: 'Your Mental Health Ally').

This manipulation also worked as a control question to assess participants' attention. Following the presentation of one of the above descriptions (as per our between-subject design), we asked

participants to answer who the chatbot was designed for: Customer support, Unemployed, Adults, or Teens.

5 STUDY 3: MAIN STUDY

Following our description of our independent variables, we outline the setup, implementation, and included measures for the main study. We set up a 2×2 between groups design to assess the impact of the two independent variables *PMA* (Low/High) and *target audience* (Teen/Adult) on three dependent variables: trust, likeability, and perceived safety (see Section 5.3).

5.1 Participants

We calculated the required sample size through a power analysis using G*Power 3.1 [29] to minimise type II errors. We use medium-to-large effect size (f^2 = 0.2), an alpha level of 0.05, and a power of 0.8, suggesting a sample size of N = 280 [53]. Using Prolific, we recruited a balanced sample of 280 participants (134 female, 141 male, 5 anonymous, average age: 37.2, SD: 12.1) from the UK and the US. We excluded one participant due to providing low effort ratings and text input (final N = 279). Participants could participate in the study using any desktop device. Recruitment parameters were set to participants having a minimum number of 100 previous submissions and a 95% approval rate. Participants who took part in pre-studies could not participate in the main study. On average, the study took the participants about ten minutes to complete. Participants were compensated using an hourly rate of £9.00. Before starting the study, participants received information on the purpose of the study and then gave informed consent.

5.2 Chatbot Interaction

We used the Botstar platform to design two interactive chatbot prototypes (see Appendix C). Participants interacted with one of the study chatbots based on their assigned condition. Participants made use of multiple-choice buttons to provide their responses to the chatbot. This design was inspired by the Woebot design (see Figure 1), and allowed us to constrict the possible directions of conversation – avoiding undesired or dangerous advice or comments (e.g., from open-ended conversation supported through a large language model). Across all four conditions, participants were presented with an equal number of 11 conversational turns (e.g., 2 conversational turns: 1) "Ugh, ever had one of those days where you just can't get work done?", 2) "All the time".) For example, following the initiation of interaction for the PMA High condition "Ugh, ever had one of those days where you just can't get work done?", participants had three options: 'All the time', 'Sometimes', and 'Never'. Choosing either of the first two options generated "Yeah, me too! Procrastination can be a real problem", whereas the last option resulted in "Good for you! You represent the .0001 percent of people who don't procrastinate!".

5.3 Experimental measures

Directly following the interaction with the chatbot, participants provided ratings on three different scales: the Multi-Dimensional Measure of Trust (MDMT) [83], Godspeed III, and Godspeed V [7], as based on our feasibility study and recent research [2, 8, 82, 93]. These scales are typically deployed when assessing influential factors on robots or computers [70], and are commonly used in HCI and HRI [87].

MDMT is a tool based on human-human/machine interaction research [83]. It has transformed findings into a set of items validated through analysis. The scale consists of four components with five items each, presented as an 8-point discrete rating scale. The four components are Capable, Ethical, Sincere, and Reliable. For example, the Capable component includes Accurate, Rigorous,

and Diligent items. Another example is Sincere, which includes Genuine, Truthful, and Authentic items.

The Godspeed questionnaire is a validated scale designed as a tool for researchers and developers in Human-Robot interaction to assess concepts related to a higher quality of interaction [7]. The 5-point Likert scale consists of five components: (I) Anthropomorphism, (II) Animacy, (III) Likeability, (IV) Perceived Intelligence, and (V) Perceived Safety. For example, the Anthropomorphism questionnaire contains five items (e.g., Fake–Natural, Artificial–Lifelike, and Machinelike–Humanlike). The questionnaire aims to capture people's perceptions of robots. However, studies have assessed people's perceptions of machines not usually considered embodied robots [6]. Following our feasibility study, we included Godspeed III and Godspeed V.

Additionally, we asked participants to answer two open questions to allow them to elaborate by expressing ideas, opinions, or insights more in-depth:

1: How do your expectations for a mental health chatbot differ from different types of chatbots (for example, customer support)?
2: What, if any, differences do you expect in a chatbot aimed at teenagers as compared to a chatbot for adults?

Following study completion, we explained the purpose of our study. We explicitly highlighted that the prototypes may be perceived as emotionally upsetting by endorsing unpopular behaviours or attitudes and that the chatbot is not deployed in any real-world settings.

5.4 Analysis

Next, we describe how we analysed the collected data using qualitative and quantitative methods.

5.4.1 Statistical testing. We conducted the statistical analysis of our quantitative data using R. We ran a 2 (PMA: Low, High) × 2 (target audience: Teen, Adult) between-groups comparison. We compared included factors (trust, likeability, and perceived safety) between conditions by conducting three two-way ANOVAs. To ensure an equal distribution of conditions, we assessed the homogeneity of variances by running Bartlett's tests for each variance test. Moreover, we calculated Cronbach's alpha to control for internal consistency.

5.4.2 Thematic analysis. We analysed our qualitative data using reflexive thematic analysis, following the six-step iterative process as suggested by Braun et al. [12]. First, we familiarised ourselves with the data by reading all of the 558 responses submitted by our participants (two open questions per participant). Following this, we systematically coded all relevant responses by following the responses to either one of the two questions. The coding process followed the technique outlined by Braun et al., including the use of highlighters to indicate potential patterns in the data [12, p. 89]. This helped us systematically review the data and code for as many potential patterns as possible, decreasing the chances of overlooking anything of interest. Third, and following this first round of coding, we re-read and revised codes into a total of seven meaningful units (*Non-clinical*, *Desirable behaviours*, *Less desirable behaviours*, *Anthropomorphism*, *Communication style*, *Appearance*, and *Carefulness*). This process was completed over multiple days, further ensuring that any additional patterns of interest were captured. Fourth, guided by these seven meaningful units, we sorted and categorised the generated codes into five themes, visualising and structuring our data through the use of tables where we further iterated on the previously highlighted patterns. Following the suggestion to further review these themes [12], we collapsed two initially separate themes into one theme. Fifth, by making structural use of tables, we obtained a representative overview to define and name the themes. By returning to our codes, meaningful units, and categories of meaningful

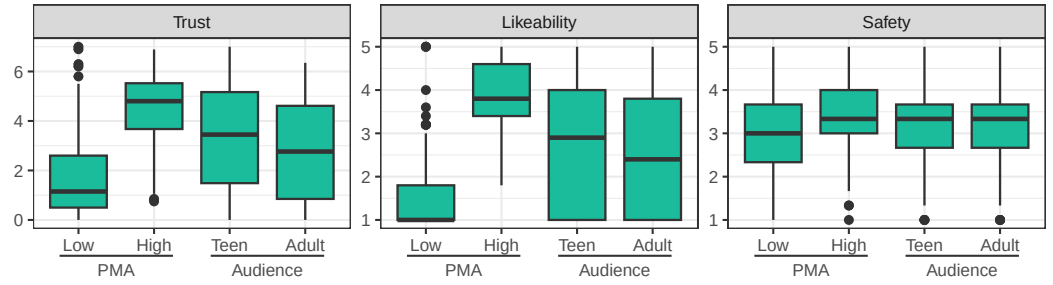

Fig. 2. Differences in trust (range 0-7), likeability (range 1-5), and perceived safety (range 1-5) between conditions (PMA: Low, High; Audience: Teen, Adult).

units, we could extract the essence of our themes, making the themes representative of the data and our research perspective (see Table 2).

We followed a straightforward structure of data extraction, similarly suggested by Braun et al. [12, p. 88]. To illustrate, the participant quote "*For it to appear real and genuine and not like I'm talking to a robot. If you're talking to yourself, it feels unreal [...]*", was coded as both 'more real' and 'more genuine', and categorised as 'Anthropomorphism'. The participant quote "*I would expect a mental health chatbot to be more compassionate, understanding, and more careful with its responses [...]*", was coded as both 'compassionate' and 'understanding', and categorised as 'Warmth'. The participant quote "*I thought that it would be more sensitive and reactive. However, unless I gave one of the prescribed answers/responses it didn't do anything [...]*", was coded as both 'sensitive' and 'reactive', and categorised as 'Sensitivity'. Lastly, the participant quote "*I would expect a chatbot for teenagers to be more relaxed, use of emojis, and shorter messages to keep engagement. Adults would typically prefer more of a human-like conversation [...]*", was coded as 'relaxed', 'emojis', 'shorter messages', and 'keep engagement', categorised as 'Appearance manifestation'. Finally, we selected and report representative quotes for each of the four themes to increase transparency and readability.

5.5 Findings

We provide an overview of how participants experienced the chatbot prototype by reporting the results from three two-way ANOVAs. We complement these results by presenting four qualitative themes resulting from our thematic analysis of open-text responses.

5.5.1 Trust, likeability, and perceived safety. For each scale, we first computed Cronbach's alpha scores to validate their internal consistency. The resulting Cronbach's alpha scores range from 0.6 to 0.9, indicating a medium to strong internal consistency. We then ran Bartlett tests of homogeneity of variances, which showed no significant differences in variance between conditions. Finally, we ran two-way ANOVAs for each of our three measures (see Table 1 and Figure 2).

For **trust**, we found a significant difference between PMA Low (M = 1.70, SD = 1.58) and PMA High (M = 4.47, SD = 1.41). Similarly, we found a significant difference between target audience, with Teenagers (M = 3.33, SD = 2.07) scoring higher than Adults (M = 2.83, SD = 1.98). We did not find an interaction effect between PMA and target audience for trust. For **likeability**, there was a significant difference, with PMA Low (M = 1.48, SD = 0.85) scoring lower than PMA High (M = 3.89, SD = 0.83). Moreover, we found a significant difference between Teenagers (M = 2.81, SD = 1.47) and Adults (M = 2.55, SD = 1.47). We did not find an interaction effect between PMA and target audience. For **perceived safety** there was a statistically significant difference in PMA

Table 1. Results of two-way ANOVAs for each of our three measures. PMA has a significant influence on all measures and the target audience significantly influences trust and likeability. We did not find an interaction effect between PMA and Target audience.

Measure	Factor	Df	F	p-value		η^2
Trust	PMA	1	244.42	< 0.001	***	0.463
	Audience	1	7.49	0.007	**	0.014
	PMA:Audience	1	0.19	0.662		< 0.001
Likeability	PMA	1	574.33	< 0.001	***	0.670
	Audience	1	6.43	0.012	*	0.008
	PMA:Audience	1	0.39	0.532		< 0.001
Perceived Safety	PMA	1	11.95	< 0.001	***	0.042
	Audience	1	0.11	0.745		< 0.001
	PMA:Audience	1	0.26	0.609		< 0.001

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

score, where participants rated perceived safety lower for PMA Low (M = 3.01, SD = 0.91) than PMA High (M = 3.37, SD = 0.80). We found no significant differences between target audience, nor an interaction effect between PMA and target audience.

Table 2. Overview of the qualitative analysis process across four steps. From left to right: We report the number of codes and include representative sample codes retrieved from our qualitative data and meaningful units subsequently mapped to the emerging four themes. 1: Anthropomorphism; 2: Non-clinical; 3: Less desirable behaviours; 4: Desirable behaviours; 5: Carefulness; 6: Appearance; 7: Communication style.

Codes (N)	Sample Codes	Units	Themes
34	More real, Less automated, More alive, Less alien, Less computer, Less fake	1	**Anthropomorphism**
130	More kindness, More caring, More friendly, More empathetic, More encouraging, More supportive	2	**Warmth**
145	More guiding, More sensitive, Less goal-oriented, More responsive, More safeguarding, More sugar-coated	3, 4, 5	**Sensitivity**
142	Slang, More relaxed, More relatable, More trendy, More emojis, More engaging	6, 7	**Appearance manifestation**

5.5.2 Qualitative results. We asked our participants two open-ended questions: 1) How do your expectations for a mental health chatbot differ from different types of chatbots (for example, customer support)?, and 2) What, if any, differences do you expect in a chatbot aimed at teenagers as compared to a chatbot for adults? We present our findings across four primary themes: Anthropomorphism,

Warmth, Sensitivity, and Appearance manifestation. In the following, we report the prevalence of the number of participants in each theme and include participant quotes to illustrate these themes.
Anthropomorphism. Several participants (N = 24) expected mental health chatbots to have different qualities than customer support chatbots. Specifically, these participants expected mental health chatbots to be more sophisticated compared to customer service chatbots, by utilising qualities crucial for human-human interactions (i.e., anthropomorphism). One participant described this as mental health chatbots displaying anthropomorphic qualities in more authentic ways:

> "I would expect more detail and authentic responses, including more questions for the person to vent. Different from the clinical, find a solution as fast as possible method customer service bots use." (P195)

Another participant explained how mental health chatbots must take different human qualities that are considered essential in sensitive contexts into consideration (e.g., caring qualities):

> "A mental health chatbot should be compassionate and interact with the talker in a human way, as mental health issues is a very sensitive and unique topic not related to other chatbots." (P197)

Human mental health professionals are trained in dealing with delicate situations to avoid harming humans. One participant highlighted how it can be frustrating to interact with customer care chatbots, and expected mental health chatbots to avoid causing humans frustration:

> "Has to be more human-like. Customer service ones can be frustrating when they can't answer your questions." (P74)

Whilst participants expected mental health chatbots to have anthropomorphic qualities, two participants also expected mental health chatbots to appear more real whilst not overplaying its behaviour:

> "I'd expect it to be less clinical and more like talking to an actual human or a friend but I think that's very hard to achieve from a chatbot. Knowing it's a bot affects the experience too." (P178)

> "Seems quite robotic when it is doing mental health as opposed to customer service. Tries too hard to be friendly." (P38)

Thus, these participants expected mental health chatbots to have more anthropomorphic qualities while avoiding coming off as trying too hard. Another participant described this more negatively, by describing how mental health chatbots cannot succeed in creating a feeling of interacting with a human:

> "I'd struggle to engage with a mental health chatbot—compared to, say, a customer support one. I dislike CS (Customer support) ones generally, but at least I'm using it for a strictly impersonal reason; I need to know something or simply use it to navigate a CS system. Using a mental health one feels alien—if I'm struggling, I'd want to think I'm at least interacting with another human being." (P86)

The above participant explicitly described the challenges of mental health chatbots, and that mental health chatbots need to be more sophisticated. In parallel, one participant described the potential benefits of mental health chatbots:

> "I found the experience strange and alien despite being familiar with using customer service chatbots. Mental health feels like it should be more personal but I can see the advantages to having 'someone' available at all times in a crisis." (P41)

Comparing human and chatbot mental health support, one participant pointed out such a challenge by describing how to expect mental health chatbots to be receptive to indirect stimuli.

> "Mental health chatbots need to be very receptive to what isn't being said as well as what is." (P198)

Warmth. Distinct but closely connected to anthropomorphism, many participants (N = 75) described warmth as something they expect in mental health chatbot behaviour. In summary, participants expected mental health chatbots to display higher warmth by being less methodical, more welcoming, and more considerate than for example customer care chatbots. Two of the participants described this as expecting mental health chatbots to be less systematic whilst displaying more warmth:

> "I would expect it to be kinder and not as methodical. I feel like people are looking to be heard with some help, not just logistical help." (P240)

> "I would expect a mental health chatbot to be more friendly and not so robotic. A mental health chatbot would need to be very sophisticated." (P264)

Participants described warmth as crucial for mental health chatbots, whilst they acknowledged the difficulties of realising more functional mental health chatbots. Moreover, two participants described how warmth is key when dealing with more vulnerable people and identify vulnerable situations to direct the person to a human professional:

> "I'd expect the chatbot to be warming, kind, considerate to human emotion—while able to provide resources to help and aid with mental health considering if a person presents certain aspects of vulnerability, it should be able to connect the person to a professional." (P262)

> "I would expect a mental health chatbot to be more compassionate, understanding and more careful with its responses. It needs to take care not to escalate a bad situation." (P109)

Another participant described attentiveness to be a key attribute of mental health chatbots, making sense of mental health chatbots by drawing similarities to puppeteering:

> "Here, the goal is *not* about finding answers. It's about listening and learning through role-play, and the simple experience of interaction. It felt akin to a Mister Rogers puppet time, redefined for the Teenage techno-sphere. I was initially unimpressed by the idea of a chatbot replacing the power of human interaction in therapy, but now I am intrigued about its potential. I think this may have merit." (P177)

Another participant made sense of this by picturing themselves as being exposed to such mental health chatbots, and expected mental health chatbots to be perspective-taking:

> "I tried to imagine myself as a teen with mild depression and stress about school work and family: what I expected from this facility was someone who is genuine, has a similar experience, is easy to chat with, is quietly confident in what they are sharing with me, and one that does not make me feel inadequate, or left out, but gives me a feeling of hope." (P191)

Moreover, two participants described how they expected mental health chatbots to be empathetic or manifest empathy, by, for example, expressing kindness:

> "I would expect a high level of empathetic values and a higher level of skill set and flexibility in chatbots dealing with mental health." (P269)

> "I would expect the language used to be more warm and welcoming, and attempt to personalise the experience a bit more." (P16)

Sensitivity. Many participants (N = 92) described that they expected differences between a mental health bot designed for adults or teenagers. Two participants described these expectations of mental health chatbots for teenagers to behave in more careful and comforting ways:

> "Teens and adults can sometimes have very different problems. Teens are so young, they have not developed any coping strategies to deal with the difficulties they face. A chatbot

> aimed at teens might be more aimed at telling teens they will be OK and they are not alone in facing their problem…" (P1)
>
> "The chatbot should be approachable and non-judgemental. It also needs to be very careful with its responses and able to improve the situation." (P109)

Two participants described their expectations of mental health chatbots to affect teens in more subtle or gentle ways, as teens are less equipped to deal with real-world problems:

> "Teens don't have as much understanding of their feelings and need subtle guidance and hand holding while making them look independent in their decisions." (P7)
>
> "The chatbot for teens might have to be more gentle in the probing nature of its questions than an adult who may be more able to deal with serious, existential issues." (P54)

Moreover, one participant expected mental health chatbots for teens to be less sincere and instead use humour as a strategy for connecting with the user:

> "I would expect a chatbot aimed at teens to be slightly more upbeat, use some humour, ask them what music they were interested in, or movies, TV and so on and get on their level." (P100)

Another participant similarly expected mental health chatbots for teenagers to avoid being confrontational and instead be relatable to teenagers' concerns:

> "I would expect it to be less confrontational, and certainly less judgemental, I would hope it used more contemporary language and more attuned to Teen's concerns on relationships, self-image, and academic pressures." (P126)

Further, two participants described how adults and teenagers face different types of problems, and how mental health chatbots should be able to address this in careful ways compared to adults:

> "Different scenarios of problems, whilst a teen may be struggling with bullying or relationship problems, an adult may be faced with a missed mortgage debt payment which can be just as taxing on a mind." (P181)
>
> "I would expect the one aimed at teens to be kinder and more sympathetic whilst also making them feel supported. I would expect the adult one to be a bit more to the point and less sugar-coated." (P30)

Appearance manifestation. Many participants (N = 88) described chatbot appearance (i.e., how the chatbot came off) as a key difference between mental health chatbots and customer support chatbots. Two participants explicitly described this as mental health chatbots for teens to manifest/come off as having certain types of characteristics:

> "It would hopefully be more understanding, and kind in its suggestions. I think the issue with the chatbot is that it didn't really care about the reasons for procrastination, but often they can be a lot more in-depth than being lazy. Teens would be more vulnerable to this kind of criticism." (P130)
>
> "I would expect a chatbot aimed at teens to be aware of things that are more likely to stress them over adults, such as school, their relationships with their friends and parents and online activities, such as the use of social media. I'd also expect the chatbot to respond differently in its interactions with them, such as encouraging them to talk about their problems with their parents, and the chatbot may rely on memes, acronyms or common humorous phrases to appear more relatable." (P107)

Another participant emphasised that a mental health chatbot for teenagers should try to express itself more directly to a teen audience, by behaving in ways appealing to teens:

> "I think the chatbot would have to 'talk' in the manner teens do. I think the example used language that would appeal to teens, it wasn't patronising but was friendly and engaging. Also, I liked that it recognised when to stop talking. From what I remember of being a Teen, I'd get bored and leave if I was being drawn into a long conversation." (P202)

Moreover, eight participants described the importance of appropriate use of language. Of these, two participants emphasised the importance of avoiding coming off as someone trying too hard:

> "Maybe slightly different language, but without trying too hard to be cool." (P41)

> "Nothing, don't make it sound 'hip' or 'cool' to try to connect with a Teen. I have twin teens and they would think it is horrible and just roll their eyes. If you are going to make a chatbot to talk to a teen just be real, don't fake it." (P252)

Another participant described different use of language as an expected difference for mental health chatbots targeted towards teens and adults. More specifically, they described different use of language as strategic ways of communicating, for example, to be accessible for teenagers:

> "I expect its language to be modelled slightly differently, perhaps approaching things from a different angle. Teens often have different mental states and emotional expectations to that of adults and a chatbot designed to help them would need to take this into account when interacting with them." (P133)

One of the participants described that they expected more casual and less complicated use of language, and to use more relevant language:

> "Use less complicated terminology and understand teenage years are difficult, the problems that may overwhelm a teen may seem small to an adult." (P78)

Two other participants also explicitly described how they expect mental health chatbots for teenagers to increase acceptance strategically, by communicating in specific ways, for example by using slang:

> "For teens, I'd perhaps expect it to use emojis or slang phrases to feel relatable and not just a bot." (P91)

> "The language the chatbot use—that is more 'street talk'." (P172)

Lastly, two participants described how they expected chatbots to communicate in more appealing ways, by using more up-to-date terminologies:

> "I think they should use words and phrases that appeal to teens." (P234)

> "Type of language used and understanding modern terminology which is constantly changing." (P104)

6 DISCUSSION

The use of chatbots to support mental health is rapidly growing, especially among teenagers [60]. Alarmingly, however, Romael Haque et al. suggest that while chatbots offer great potential for mental health support, they largely fail to provide risk-free interactions which may subsequently lead to inadequate support during times of crisis [38]. Given that those in vulnerable positions use these chatbots, it is critical to consider any potential pitfalls in using these systems. Our results indicate that perceived moral agency plays a key role in influencing perceptions towards such mental health chatbots. Based on these results, we argue that a better understanding of perceived moral agency in the context of chatbots can inform the design of more supportive chatbots in sensitive application domains.

6.1 Trust, likeability, and perceived safety

Considering our first research question, our results highlight significant differences in the trust provided towards mental health chatbots as dependent on their PMA level. This aligns with earlier approaches investigating trust and agency [28, 33, 67], which show that agency facilitates, impacts, and changes trust in robots or machines. Moreover, our results indicate that mental health chatbots are rated significantly higher in likeability and perceived safety when displaying higher PMA.

Trust has become an increasingly investigated topic in HCI [85], including in the domain of mental health support [11]. In our study, we find that peoples' trust in chatbots is higher when the chatbot displays high PMA. As such, people's perception of a chatbot's moral agency is key to increasing perceived trustworthiness, presumably indispensable for mental health chatbots designed for young people usually reluctant to use traditional mental health services [24].

In addition to trust, prior work has highlighted likeability as crucial for effective human-robot interactions [93]. Our quantitative results indicate that participants' perception of mental health chatbots improves when displaying high PMA. However, the likeability of chatbots has received limited attention in the context of mental health chatbots.

Lastly, considering perceived safety, Abd-Alrazaq et al. described that only two prior studies have assessed safety in mental health chatbots [1]. Whilst those two studies indicated no harm or adverse events caused by interacting with the chatbot, such as worsening of depressive symptoms, safety in mental health chatbots is critical. As with human-human interaction, safety is key in human-computer interactions when considering sensitive settings [21]. Our quantitative results indicate that participants rate mental health chatbots as safer when displaying high perceived moral agency. This aligns well with recent research in HRI that suggests multiple factors play a crucial role in perceived safety (e.g., comfort and familiarity [2]).

Our thematic analysis reveals themes around people's perceptions and expectations of mental health chatbots. Participants highlighted the importance of qualities related to, for example, warmth and sensitivity. These findings stand in contrast with a recent study on care robots imitating human appearance and behaviour, showing competence as more important than warmth for non-humanoids in a task-specific setting [44]. Crucially, Jung et al.'s study focused on care robots avoiding stereotypical appearance or behaviour, whereas our focus on mental health chatbots may result in different requirements due to chatbots being heavily dependent on natural language capabilities. Based on our results, we argue that mental health chatbots require different qualities than non-humanoids situated in health and well-being settings (see Section 6.2). This aligns with You et al's recent investigation on how chatbot-based symptom checkers should respond to users, where their results partly point to the value of going beyond the design of 'doctor-like' agent behaviour [91]. The implementation of such qualities can be informed by perceived moral agency to improve, for example, meaningful interactions [6] for long-term human-chatbot relationships [79].

We utilised topics and descriptions from a state-of-the-art chatbot that provided a relatively simple interaction to participants. Despite this simplicity, our results indicate that our participants perceived human qualities (i.e., moral agency) in non-human actors. This aligns with work on how people try to make sense of algorithmic systems [31, 55, 84], for example, by using mental representation [16], folk theories [23], or algorithmic personas [90]. The results presented in this study contribute to the domain of HCI and health and well-being. Our results further highlight a gap in research on how different levels of perceived moral agency can influence people's perceptions. Research on human-human interaction in cognitive science and psychology has provided strong evidence that morality changes how we perceive others [22, 54], thus pointing to morality as a relevant factor in our interactions with digital actors.

6.2 Design recommendations for mental health chatbots

Recent technological advancements in large language models have enabled users to customise the style of responses they receive in text-based interfaces (e.g., Bing) [27]. By giving users the option to calibrate a model's behaviour, its responses are more likely to align with personal preferences. New calibration options for users also raise novel challenges for chatbot designers. Not only must chatbot designers consider what options are for users to control and which are predetermined, they must also carefully define appropriate boundaries within which users can customise the model behaviour. This is a largely unexplored area featuring not only technological but also sociological challenges, as indicated above. We further elaborate on the various aspects in which users can calibrate the behaviour of mental health chatbots in the subsequent design recommendations.

Rapp et al.'s recent systematic literature review on chatbots calls attention to the importance of emotions and humanness in chatbots [72]. Our results also support the identified importance of human qualities in chatbots, indicating that people expect chatbots to carry distinct qualities as dependent on their context of use. Whilst the following design recommendations for mental health chatbots are rooted in increased 'human-ness', we also note that prior work gives reasons to be cautious of designing systems which resemble human behaviour [36]. For example, prior work shows that such humanness can cause negative feelings of eeriness [19] or creepiness [88]. These perceptions are critical to avoid in sensitive settings. However, and as our results indicate, there are domains which might benefit from a more 'humane' technology experience (e.g., compassion)—with mental health a prime example of such a domain.

Through the use of generative AI, our conversations and interactions with digital technology can appear increasingly human-like. Employing these possibilities for the better requires a careful understanding of the target audience. Considering our second research question and based on our results, we outline four design recommendations (DRs) for future work in the domain of mental health chatbot design.

DR1: Avoid overly friendly behaviour. Our results show that participants expect mental health chatbots to display increased anthropomorphism as compared to, for example, customer care chatbots. While being friendly is critical in sensitive settings, prior work points to unconditional positivity as having potentially negative consequences, as people's positive perception of other's sociability depends on perceived morality (e.g., if others are perceived as immoral, people prefer them to show less positive sociability [54]). We, therefore, recommend that mental health chatbots ask compassionate questions whilst avoiding being too friendly. This aligns well with recent findings in digital mental health interventions that emphasise individuals' changing needs as a critical factor to consider when designing for young people [64], as some may require less overly friendly behaviour.

DR2: Increase personal perspective-taking. Participants report that they expect mental health chatbots to display increased warmth compared to other types of chatbots. Displays of attentiveness and perspective-taking by communicating in kind ways are therefore critical and can, for example, be achieved by pointing out shared experiences. Prior work points to AI being perceived as less warm than people [92], yet warmth is described as critical for human-human interactions [42]. Consequently, we recommend increasing the warmth of mental health chatbots through perspective-taking. This aligns with Meyerhoff et al.'s suggestion of considering young users' ambivalence in interacting with mental health chatbots [64].

DR3: Avoid confrontation. Our participants expressed an expectation of sensitivity of mental health chatbots. Consequently, we urge designers to limit mental health chatbots' use of confrontational or demanding expressions and instead offer more positive reassurances and comforting words. Prior work identifies comfort as a key factor influencing perceived safety [2], allowing for a

further increase in sensitivity in mental health chatbots. As Meyerhoff et al. suggest, users vary in readiness for digital mental health interventions and recommend increasing young people's control over directness in their interactions with mental health chatbots [64]. Garrido et al. highlight that increasing relevance and appeal for young people is key for successful digital mental health interventions, outlining that non-confronting interventions are a promising way to increase the appeal of such interventions [34].

DR4: Control manifestations. Our results show that participants expect mental health chatbots to adapt their appearance manifestations to their audience, for example, by expressing teen-dependent language to avoid appearing as 'corny'. Based on our data, we recommend designing mental health chatbots that use target audience-dependent language to indicate awareness of relevant target audience-dependent topics without overdoing it. Prior work on digital agents showed that coming off as trying too hard results in reduced trust [20]. Chatbots' controlling manifestations can thus be related to the idea of self-presentation, which essentially is about engineering one's attributes to control others' perception of oneself [81]. Our results also suggest that participants had specific expectations of the communication style of mental health chatbots. Practically, this could be achieved by using appealing and up-to-date terminology and emoji, as noted by prior work on the role of language strategies [17] and conversational style [48].

6.3 Limitations & Future Work

We recognise several limitations in this work, both related to the application domain and the chosen participant sample.

Due to various contextual and subjective factors (e.g., cultural and environmental factors), measuring perceived moral agency in mental health chatbots is highly challenging [3]. However, recent work emphasises the importance of better understanding morality as a continuous and dynamic concept, rather than steady and constant [86]. By using established measures immediately after the interaction, chances are increased to capture morality in a mental health context. As morality depends on numerous factors, including its interchangeable nature, limiting the cultural diversity of the included sample was necessary. Despite these limitations, real-world interventions could be an interesting next step as we still rely on self-report responses. Indeed, to see any real impact from mental health chatbots being deployed in real-world scenarios, longitudinal evaluations are necessary to better understand any potential lasting effects.

The topic that participants discussed with the chatbots, procrastination, only represents a small aspect of topics in the mental health domain. Therefore, it is necessary to further explore other topics within mental health (e.g., depression [15, 62, 69]). As prior work highlights that systems can cause harm when interacting with people in vulnerable situations [9], a careful design and research approach is required. In light of these ethical concerns, we chose to present participants with a less intrusive topic. While we expect our quantitative results to replicate even stronger when presented with more severe mental health topics, different perspectives may emerge among participant preferences for the design of chatbots.

We limited our participant sample to UK and US participants to ensure a sufficient understanding of English. However, this limits our results to their perspectives, even though most people are not 'WEIRD' [40]. Our sample consisted of adults with an average age of 37 years. Indeed, this has an impact on the generalisability of our results, as the average age of 37 does not represent younger age groups. However, we see the relevance of including ratings and perspectives from an older age group, as those are potentially more experienced in dealing with challenges younger people might face. Future studies might want to investigate the specific needs of a teenage audience — a challenging task that can be done with people situated in more sensitive settings (e.g., teenagers) through participatory design approaches (e.g., with vulnerable groups [25]). Similarly, the two

conditions in our study manipulated the target audience — a more generic bot with no target demographics could be useful to include in future investigations.

Lastly, our study was limited to a relatively straightforward mental health chatbot in which interaction is structured through multiple-choice options. As such, our prototype does not significantly differ in functionality as compared to state-of-the-art mental health chatbots, which are similarly constrained in the interaction provided (e.g., Woebot or Wysa). This constraint is likely put in place to ensure that no harmful or dangerous messages are accidentally sent to the user.

Considering people's tendency to anthropomorphise non-humans [66], there is a need to further explore people's perception of moral agency when faced with more advanced chatbots introduced to different contexts (e.g., mild, moderate, and severe disorders). A particularly interesting way forward is to assess people's perception of moral agency in conversational models able to technically provide and realise 'empathetic' responses [61, 73, 76], since implementations of LLM-based applications are now being explored in different domains (e.g., in medical sciences [59, 80]). We therefore call on the broader HCI research community to consider peoples' perceptions of system morality as a worthwhile avenue for future work.

7 CONCLUSION

Chatbots are increasingly introduced in sensitive settings, such as mental health for young people. This is done without fully understanding how people make sense of chatbots, and how such chatbots may impact people—either positively and negatively. In this paper, we showed that by manipulating the PMA of a chatbot, we can influence key factors critical to human-chatbot interactions such as trust, likeability, and perceived safety. This shows that PMA plays a crucial role in mental health chatbots, which are typically used by users in a vulnerable position. By manipulating PMA we can optimise these factors to meet peoples' expectations of mental health chatbots. We have provided design recommendations to meet peoples' expectations of mental health chatbots. Future research should extend this work on how people make sense of mental health chatbots, and how PMA influences human-computer interactions at large.

ACKNOWLEDGMENTS

This work is supported by the Carlsberg Foundation, grant CF21-0159.

REFERENCES

[1] Alaa Ali Abd-Alrazaq, Asma Rababeh, Mohannad Alajlani, Bridgette M Bewick, and Mowafa Househ. 2020. Effectiveness and Safety of Using Chatbots to Improve Mental Health: Systematic Review and Meta-Analysis. *Journal of Medical Internet Research* 22, 7 (2020), e16021. https://doi.org/10.2196/16021

[2] Neziha Akalin, Annica Kristoffersson, and Amy Loutfi. 2022. Do you feel safe with your robot? Factors influencing perceived safety in human-robot interaction based on subjective and objective measures. *International Journal of Human-Computer Studies* 158 (2022), 102744. https://doi.org/10.1016/j.ijhcs.2021.102744

[3] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The Moral Machine Experiment. *Nature* 563 (2018). https://doi.org/10.1038/s41586-018-0637-6

[4] Edmond Awad, Sydney Levine, Michael Anderson, Susan Leigh Anderson, Vincent Conitzer, M.J. Crockett, Jim A.C. Everett, Theodoros Evgeniou, Alison Gopnik, Julian C. Jamison, Tae Wan Kim, S. Matthew Liao, Michelle N. Meyer, John Mikhail, Kweku Opoku-Agyemang, Jana Schaich Borg, Juliana Schroeder, Walter Sinnott-Armstrong, Marija Slavkovik, and Josh B. Tenenbaum. 2022. Computational ethics. *Trends in Cognitive Sciences* 26, 5 (2022), 388–405. https://doi.org/10.1016/j.tics.2022.02.009

[5] Petter Bae Brandtzæg, Marita Skjuve, Kim Kristoffer Kristoffer Dysthe, and Asbjørn Følstad. 2021. When the Social Becomes Non-Human: Young People's Perception of Social Support in Chatbots. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 257, 13 pages. https://doi.org/10.1145/3411764.3445318

[6] Jaime Banks. 2019. A perceived moral agency scale: Development and validation of a metric for humans and social machines. *Computers in Human Behavior* 90 (2019), 363–371. https://doi.org/10.1016/j.chb.2018.08.028

[7] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics* 1, 1 (2009), 71–81. https://doi.org/10.1007/s12369-008-0001-3

[8] Timothy Bickmore and Amanda Gruber. 2010. Relational Agents in Clinical Psychiatry. *Harvard Review of Psychiatry* 18, 2 (2010), 119–130. https://doi.org/10.3109/10673221003707538

[9] Douglas Birsch. 2004. Moral Responsibility for Harm Caused by Computer System Failures. *Ethics and Information Technology* 6 (2004), 233–245. https://doi.org/10.1007/s10676-005-5609-5

[10] Paula Boddington. 2020. AI and moral thinking: how can we live well with machines to enhance our moral agency? *AI and Ethics* 1 (2020). https://doi.org/10.1007/s43681-020-00017-0

[11] Kyle Boyd, Courtney Potts, Raymond Bond, Maurice Mulvenna, Thomas Broderick, Con Burns, Andrea Bickerdike, Mike Mctear, Catrine Kostenius, Alex Vakaloudis, Indika Dhanapala, Edel Ennis, and Fred Booth. 2022. Usability Testing and Trust Analysis of a Mental Health and Wellbeing Chatbot. In *Proceedings of the 33rd European Conference on Cognitive Ergonomics* (Kaiserslautern, Germany) *(ECCE '22)*. Association for Computing Machinery, New York, NY, USA, Article 18, 8 pages. https://doi.org/10.1145/3552327.3552348

[12] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[13] Sandra Bucci, Matthias Schwannauer, and Natalie Berry. 2019. The digital revolution and its impact on mental health care. *Psychology and Psychotherapy: Theory, Research and Practice* 92, 2 (2019), 277–297. https://doi.org/10.1111/papt.12222

[14] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (2021), 21 pages. https://doi.org/10.1145/3449287

[15] Eleanor R. Burgess, Madhu C. Reddy, and David C. Mohr. 2022. "I Just Can't Help But Smile Sometimes": Collaborative Self-Management of Depression. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 70 (2022), 32 pages. https://doi.org/10.1145/3512917

[16] John M. Carroll and Judith Reitman Olson. 1988. Mental Models in Human-Computer Interaction. In *Handbook of Human-Computer Interaction*, Martin Helander (Ed.). North-Holland, Amsterdam, 45–65. https://doi.org/10.1016/B978-0-444-70536-5.50007-5

[17] Justine Cassell and Timothy Bickmore. 2001. Negotiated Collusion: Modeling Social Language and its Relationship Effects in Intelligent Agents. *User Modelling and User-Adapted Interaction* 13 (2001). https://doi.org/10.1023/A:1024026532471

[18] Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How Should My Chatbot Interact? A Survey on Social Characteristics in Human–Chatbot Interaction Design. *International Journal of Human–Computer Interaction* 37, 8 (2021), 729–758. https://doi.org/10.1080/10447318.2020.1841438

[19] Leon Ciechanowski, Aleksandra Przegalinska, Mikolaj Magnuski, and Peter Gloor. 2019. In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems* 92 (2019), 539–548. https://doi.org/10.1016/j.future.2018.01.055

[20] Henriette Cramer, Vanessa Evers, Tim van Slooten, Mattijs Ghijsen, and Bob Wielinga. 2010. Trying Too Hard: Effects of Mobile Agents' (Inappropriate) Social Expressiveness on Trust, Affect and Compliance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1471–1474. https://doi.org/10.1145/1753326.1753546

[21] Munmun De Choudhury and Emre Kiciman. 2018. Integrating Artificial and Human Intelligence in Complex, Sensitive Problem Domains: Experiences from Mental Health. *AI Magazine* 39, 3 (2018), 69–80. https://doi.org/10.1609/aimag.v39i3.2815

[22] Jean Decety and Jason M. Cowell. 2014. The complex relation between morality and empathy. *Trends in Cognitive Sciences* 18, 7 (2014), 337–339. https://doi.org/10.1016/j.tics.2014.04.008

[23] Michael A. DeVito, Darren Gergle, and Jeremy Birnholtz. 2017. "Algorithms Ruin Everything": #RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3163–3174. https://doi.org/10.1145/3025453.3025659

[24] Gilly Dosovitsky and Eduardo Bunge. 2022. Development of a chatbot for depression: adolescent perceptions and recommendations. *Child and Adolescent Mental Health* (2022). https://doi.org/10.1111/camh.12627

[25] Ana Maria Bustamante Duarte, Nina Brendel, Auriol Degbelo, and Christian Kray. 2018. Participatory Design and Participatory Research: An HCI Case Study with Young Forced Migrants. *ACM Trans. Comput.-Hum. Interact.* 25, 1, Article 3 (feb 2018), 39 pages. https://doi.org/10.1145/3145472

[26] Tessa Eagle, Aman Mehrotra, Aayush Sharma, Alex Zuniga, and Steve Whittaker. 2022. "Money Doesn't Buy You Happiness": Negative Consequences of Using the Freemium Model for Mental Health Apps. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 265 (2022), 38 pages. https://doi.org/10.1145/3555155

[27] Benj Edwards. 2023. AI-powered Bing Chat gains three distinct personalities. *Ars Technica* (2023). https://arstechnica.com/information-technology/2023/03/microsoft-equips-bing-chat-with-multiple-personalities-creative-balanced-precise/

[28] Vegard Engen, J. Brian Pickering, and Paul Walland. 2016. Machine Agency in Human-Machine Networks; Impacts and Trust Implications. In *Human-Computer Interaction. Novel User Experiences*, Masaaki Kurosu (Ed.). Springer International Publishing, Cham, 96–106. https://doi.org/10.1007/978-3-319-39513-5_9

[29] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical Power Analyses Using G*Power 3.1: Tests for Correlation and Regression Analyses. *Behavior research methods* 41 (2009), 1149–60. https://doi.org/10.3758/BRM.41.4.1149

[30] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health* 4 (2017), e19. https://doi.org/10.2196/mental.7785

[31] Jodi Forlizzi and Katja Battarbee. 2004. Understanding Experience in Interactive Systems. In *Proceedings of the 5th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques* (Cambridge, MA, USA) *(DIS '04)*. Association for Computing Machinery, New York, NY, USA, 261–268. https://doi.org/10.1145/1013115.1013152

[32] Paul Formosa and Malcolm Ryan. 2021. Making Moral Machines: Why We Need Artificial Moral Agents. *AI Soc.* 36, 3 (2021), 839–851. https://doi.org/10.1007/s00146-020-01089-6

[33] Chelsea Frazier-Young, Malcolm McCurry, Kevin Zish, and Greg Trafton. 2022. Perceived Agency Changes Performance and Moral Trust in Robots. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 44.

[34] Sandra Garrido, Chris Millington, Daniel Cheers, Katherine Boydell, Emery Schubert, Tanya Meade, and Quang Vinh Nguyen. 2019. What Works and What Doesn't Work? A Systematic Review of Digital Mental Health Interventions for Depression and Anxiety in Young People. *Frontiers in Psychiatry* 10 (2019). https://doi.org/10.3389/fpsyt.2019.00759

[35] Gerd Gigerenzer. 2010. Moral Satisficing: Rethinking Moral Behavior as Bounded Rationality. *Topics in Cognitive Science* 2, 3 (2010), 528–554. https://doi.org/10.1111/j.1756-8765.2010.01094.x

[36] J. P. Grodniewicz and Mateusz Hohol. 2023. Waiting for a digital therapist: three challenges on the path to psychotherapy delivered by artificial intelligence. *Frontiers in Psychiatry* 14 (2023). https://doi.org/10.3389/fpsyt.2023.1190084

[37] Amanda M. Hall, Paulo H. Ferreira, Christopher G. Maher, Jane Latimer, and Manuela L. Ferreira. 2010. The Influence of the Therapist-Patient Relationship on Treatment Outcome in Physical Rehabilitation: A Systematic Review. *Physical Therapy* 90, 8 (2010), 1099–1110. https://doi.org/10.2522/ptj.20090245

[38] M D Romael Haque and Sabirat Rubya. 2023. An Overview of Chatbot-Based Mobile Mental Health Apps: Insights From App Description and User Reviews. *JMIR Mhealth Uhealth* 11 (2023), e44838. https://doi.org/10.2196/44838

[39] Md Romael Haque and Sabirat Rubya. 2022. "For an App Supposed to Make Its Users Feel Better, It Sure is a Joke" - An Analysis of User Reviews of Mobile Mental Health Applications. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 421 (2022), 29 pages. https://doi.org/10.1145/3555146

[40] Joseph Henrich, Steven Heine, and Ara Norenzayan. 2010. Most people are not WEIRD. *Nature* 466 (2010), 29. https://doi.org/10.1038/466029a

[41] Annabell Ho, Jeff Hancock, and Adam Miner. 2018. Psychological, Relational, and Emotional Effects of Self-Disclosure After Conversations With a Chatbot. *Journal of Communication* 68 (2018). https://doi.org/10.1093/joc/jqy026

[42] Lauren C. Howe, Kari A. Leibowitz, and Alia J. Crum. 2019. When Your Doctor "Gets It" and "Gets You": The Critical Role of Competence and Warmth in the Patient–Provider Interaction. *Frontiers in Psychiatry* 10 (2019). https://doi.org/10.3389/fpsyt.2019.00475

[43] Ryan Blake Jackson and Tom Williams. 2021. A Theory of Social Agency for Human-Robot Interaction. *Frontiers in Robotics and AI* 8 (2021). https://doi.org/10.3389/frobt.2021.687726

[44] Frederike Jung, Heiko Müller, and Susanne CJ Boll. 2022. It's Not Warm But That's Okay: About Robots That Avoid Human Stereotypes. In *Nordic Human-Computer Interaction Conference* (Aarhus, Denmark) *(NordiCHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 48, 15 pages. https://doi.org/10.1145/3546155.3546695

[45] Takeshi Kamita, Atsuko Matsumoto, Boyu Sun, and Tomoo Inoue. 2020. Promotion of Continuous Use of a Self-Guided Mental Healthcare System by a Chatbot. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing* (Virtual Event, USA) *(CSCW '20 Companion)*. Association for Computing Machinery, New York, NY, USA, 293–298. https://doi.org/10.1145/3406865.3418343

[46] Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T. Hancock, and Michael S. Bernstein. 2020. Conceptual Metaphors Impact Perceptions of Human-AI Collaboration. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 163 (2020), 26 pages. https://doi.org/10.1145/3415234

[47] Prabha Khannan. 2022. Is It My Turn Yet? Teaching a Voice Assistant When to Speak. *HAI Stanford* (2022). https://hai.stanford.edu/news/it-my-turn-yet-teaching-voice-assistant-when-speak

[48] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing Data from Chatbot and Web Surveys: Effects of Platform and Conversational Style on Survey Response Quality. In *Proceedings of the 2019 CHI Conference on Human*

Factors in Computing Systems (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300316

[49] Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. 2018. Reflection Companion: A Conversational System for Engaging Users in Reflection on Physical Activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2 (2018), 1–26. https://doi.org/10.1145/3214273

[50] Thomas Kosch, Robin Welsch, Lewis Chuang, and Albrecht Schmidt. 2022. The Placebo Effect of Artificial Intelligence in Human-Computer Interaction. *ACM Trans. Comput.-Hum. Interact.* (2022). https://doi.org/10.1145/3529225

[51] Theodora Koulouri, Robert D. Macredie, and David Olakitan. 2022. Chatbots to Support Young Adults' Mental Health: An Exploratory Study of Acceptability. *ACM Trans. Interact. Intell. Syst.* 12, 2, Article 11 (2022), 39 pages. https://doi.org/10.1145/3485874

[52] Michael Laakasuo, Jussi Palomäki, Anton Kunnari, Sanna Rauhala, Marianna Drosinou, Juho Halonen, Noora Lehtonen, Mika Koverola, Marko Repo, Jukka Sundvall, Aku Visala, and Kathryn B. Francis. 2022. Moral psychology of nursing robots: Exploring the role of robots in dilemmas of patient autonomy. *European Journal of Social Psychology* (2022). https://doi.org/10.1002/ejsp.2890

[53] Daniel Lakens. 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology* 4 (2013). https://doi.org/10.3389/fpsyg.2013.00863

[54] Justin F. Landy, Jared Piazza, and Geoffrey P. Goodwin. 2016. When It's Bad to Be Friendly and Smart: The Desirability of Sociability and Competence Depends on Morality. *Personality and Social Psychology Bulletin* 42, 9 (2016), 1272–1290. https://doi.org/10.1177/0146167216655984

[55] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684. https://doi.org/10.1177/2053951718756684

[56] Yi-Chieh Lee, Naomi Yamashita, and Yun Huang. 2020. Designing a Chatbot as a Mediator for Promoting Deep Self-Disclosure to a Real Mental Health Professional. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 31 (2020), 27 pages. https://doi.org/10.1145/3392836

[57] Susanna Lehtimaki, Jana Martic, Brian Wahl, Katherine T Foster, and Nina Schwalbe. 2021. Evidence on Digital Mental Health Interventions for Adolescents and Young People: Systematic Overview. *JMIR Ment Health* 8, 4 (2021), e25847. https://doi.org/10.2196/25847

[58] Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. 2020. CAiRE: An End-to-End Empathetic Chatbot. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 09 (2020), 13622–13623. https://doi.org/10.1609/aaai.v34i09.7098

[59] Siru Liu, Allison B. McCoy, Aileen P. Wright, Babatunde Carew, Julian Z. Genkins, Sean S. Huang, Josh F. Peterson, Bryan Steitz, and Adam Wright. 2023. Leveraging Large Language Models for Generating Responses to Patient Messages. *medRxiv* (2023). https://doi.org/10.1101/2023.07.14.23292669

[60] Cora Lydon. 2022. Wysa AI-chatbot app to be rolled out to teenagers across west London. *digitalhealth* (2022). https://www.digitalhealth.net/2022/09/wysa-ai-chatbot-teens-west-london/

[61] Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking Emotions for Empathetic Response Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 8968–8979. https://doi.org/10.18653/v1/2020.emnlp-main.721

[62] Laura Martinengo, Elaine Lum, and Josip Car. 2022. Evaluation of chatbot-delivered interventions for self-management of depression: Content analysis. *Journal of Affective Disorders* 319 (2022), 598–607. https://doi.org/10.1016/j.jad.2022.09.028

[63] Andreia Martins Martinho, Adam Poulsen, Maarten Kroesen, and Caspar Chorus. 2021. Perspectives about artificial moral agents. *AI and Ethics* 1 (2021). https://doi.org/10.1007/s43681-021-00055-2

[64] Jonah Meyerhoff, Rachel Kornfield, David C. Mohr, and Madhu Reddy. 2022. Meeting Young Adults' Social Support Needs across the Health Behavior Change Journey: Implications for Digital Mental Health Tools. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 312 (2022), 33 pages. https://doi.org/10.1145/3555203

[65] Joonas Moilanen, Niels van Berkel, Aku Visuri, Ujwal Gadiraju, Willem van der Maden, and Simo Hosio. 2023. Supporting mental health self-care discovery through a chatbot. *Frontiers in Digital Health* 5 (2023). https://doi.org/10.3389/fdgth.2023.1034724

[66] Andreea Muresan and Henning Pohl. 2019. Chats with Bots: Balancing Imitation and Engagement. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3313084

[67] Sari R.R. Nijssen, Barbara C. N. Müller, Tibor Bosse, and Markus Paulus. 2022. Can you count on a calculator? The role of agency and affect in judgments of robots as moral agents. *Human–Computer Interaction* 0, 0 (2022), 1–17. https://doi.org/10.1080/07370024.2022.2080552

[68] Shaunagh O'Sullivan, Niels van Berkel, Vassilis Kostakos, Lianne Schmaal, Simon D'Alfonso, Lee Valentine, Sarah Bendall, Barnaby Nelson, John F Gleeson, and Mario Alvarez-Jimenez. 2023. Understanding What Drives Long-term Engagement in Digital Mental Health Interventions: Secondary Causal Analysis of the Relationship Between Social Networking and Therapy Engagement. *JMIR Ment Health* 10 (2023), e44812. https://doi.org/10.2196/44812

[69] Wenjing Pan, Bo Feng, V. Skye Wingate, and Siyue Li. 2020. What to Say When Seeking Support Online: A Comparison Among Different Levels of Self-Disclosure. *Frontiers in Psychology* 11 (2020). https://doi.org/10.3389/fpsyg.2020.00978

[70] Jason E. Plaks, Laura Bustos Rodriguez, and Reem Ayad. 2022. Identifying psychological features of robots that encourage and discourage trust. *Computers in Human Behavior* 134 (2022), 107301. https://doi.org/10.1016/j.chb.2022.107301

[71] Claudette Pretorius, Darragh McCashin, and David Coyle. 2022. Supporting personal preferences and different levels of need in online help-seeking: a comparative study of help-seeking technologies for mental health. *Human–Computer Interaction* 0, 0 (2022), 1–22. https://doi.org/10.1080/07370024.2022.2077733

[72] Amon Rapp, Lorenzo Curti, and Arianna Boldi. 2021. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies* 151 (2021), 102630. https://doi.org/10.1016/j.ijhcs.2021.102630

[73] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 5370–5381. https://doi.org/10.18653/v1/P19-1534

[74] Chelsea Schein and Kurt Gray. 2018. The Theory of Dyadic Morality: Reinventing Moral Judgment by Redefining Harm. *Personality and Social Psychology Review* 22, 1 (2018), 32–70. https://doi.org/10.1177/1088868317698288

[75] Jen Semler. 2022. Artificial Quasi Moral Agency. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) *(AIES '22)*. Association for Computing Machinery, New York, NY, USA, 913. https://doi.org/10.1145/3514094.3539549

[76] Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021. Towards Facilitating Empathic Conversations in Online Mental Health Support: A Reinforcement Learning Approach. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 194–205. https://doi.org/10.1145/3442381.3450097

[77] Sharpbrains. 2022. AI-enabled chatbot Wysa receives FDA Breakthrough Device designation for patients with chronic pain, depression and anxiety. *Sharpbrains* (2022). https://sharpbrains.com/blog/2022/06/06/ai-enabled-chatbot-wysa-receives-fda-breakthrough-device-designation-for-patients-with-chronic-pain-depression-and-anxiety

[78] Geovana Ramos Sousa Silva and Edna Dias Canedo. 2022. Towards User-Centric Guidelines for Chatbot Conversational Design. *International Journal of Human–Computer Interaction* 0, 0 (2022), 1–23. https://doi.org/10.1080/10447318.2022.2118244

[79] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2022. A longitudinal study of human–chatbot relationships. *International Journal of Human-Computer Studies* 168 (2022), 102903. https://doi.org/10.1016/j.ijhcs.2022.102903

[80] Vera Sorin, Danna Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, and Eyal Klang. 2023. Large Language Models (LLMs) and Empathy – A Systematic Review. *medRxiv* (2023). https://doi.org/10.1101/2023.08.07.23293769

[81] Jeff Stanley. 2023. Personality for Virtual Assistants: A Self-Presentation Approach. (2023). https://doi.org/10.5772/intechopen.1001934

[82] Moonyoung Tae and Joonhwan Lee. 2020. The Effect of Robot's Ice-Breaking Humor on Likeability and Future Contact Intentions. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 462–464. https://doi.org/10.1145/3371382.3378267

[83] Daniel Ullman and Bertram F. Malle. 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* (Chicago, IL, USA) *(HRI '18)*. Association for Computing Machinery, New York, NY, USA, 263–264. https://doi.org/10.1145/3173386.3176991

[84] Niels van Berkel, Jorge Goncalves, Danula Hettiachchi, Senuri Wijenayake, Ryan M. Kelly, and Vassilis Kostakos. 2019. Crowdsourcing Perceptions of Fair Predictors for Machine Learning: A Recidivism Case Study. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 28 (2019), 21 pages. https://doi.org/10.1145/3359130

[85] Niels van Berkel, Zhanna Sarsenbayeva, and Jorge Goncalves. 2023. The methodology of studying fairness perceptions in Artificial Intelligence: Contrasting CHI and FAccT. *International Journal of Human-Computer Studies* 170 (2023), 102954. https://doi.org/10.1016/j.ijhcs.2022.102954

[86] Niels van Berkel, Benjamin Tag, Jorge Goncalves, and Simo Hosio. 2022. Human-centred artificial intelligence: a contextual morality perspective. *Behaviour & Information Technology* 41, 3 (2022), 502–518. https://doi.org/10.1080/0144929X.2020.1818828

[87] Astrid Weiss and Christoph Bartneck. 2015. Meta analysis of the usage of the Godspeed Questionnaire Series. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 381–388. https://doi.org/10.1109/ROMAN.2015.7333568

[88] Paweł W. Woźniak, Jakob Karolus, Florian Lang, Caroline Eckerth, Johannes Schöning, Yvonne Rogers, and Jasmin Niess. 2021. Creepy Technology:What Is It and How Do You Measure It?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 719, 13 pages. https://doi.org/10.1145/3411764.3445299

[89] Jesse H. Wright and Denise Davis. 1994. The therapeutic relationship in cognitive-behavioral therapy: Patient perceptions and therapist responses. *Cognitive and Behavioral Practice* 1, 1 (1994), 25–45. https://doi.org/10.1016/S1077-7229(05)80085-9

[90] Eva Yiwei Wu, Emily Pedersen, and Niloufar Salehi. 2019. Agent, Gatekeeper, Drug Dealer: How Content Creators Craft Algorithmic Personas. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 219 (2019), 27 pages. https://doi.org/10.1145/3359321

[91] Yue You, Chun-Hua Tsai, Yao Li, Fenglong Ma, Christopher Heron, and Xinning Gui. 2023. Beyond Self-Diagnosis: How a Chatbot-Based Symptom Checker Should Respond. *ACM Trans. Comput.-Hum. Interact.* 30, 4, Article 64 (sep 2023), 44 pages. https://doi.org/10.1145/3589959

[92] Zaixuan Zhang, Zhansheng Chen, and Liying Xu. 2022. Artificial intelligence and moral dilemmas: Perception of ethical decision-making in AI. *Journal of Experimental Social Psychology* 101 (2022), 104327. https://doi.org/10.1016/j.jesp.2022.104327

[93] Vivienne Jia Zhong, Nicolas Mürset, Janine Jäger, and Theresa Schmiedel. 2022. Exploring Variables That Affect Robot Likeability. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 1140–1145. https://doi.org/10.1109/HRI53351.2022.9889602

A QUESTIONS

A.1 Perceived Moral Agency Scale [6]

Morality. Based on the dialogue presented above, please rate your level of agreement with the provided statements (7-point Likert scale).

- This chatbot has a sense for what is right and wrong.
- This chatbot can think through whether an action is moral.
- This chatbot might feel obligated to behave in a moral way.
- This chatbot is capable of being rational about good and evil.
- This chatbot behaves according to moral rules.
- This chatbot would refrain from doing what has painful repercussions.

Dependency. Based on the dialogue presented above, please rate your level of agreement with the provided statements (7-point Likert scale).

- This chatbot can only behave how it is programmed to behave.
- This chatbot's actions are the result of its programming.
- This chatbot can only do what humans tell it to do.
- This chatbot would never do anything it was not programmed to do.

A.2 Godspeed Questionnaire [7]

Likeability. Please rate your impression of the chatbot on these scales (5-point Likert scale).

- Dislike / Like
- Unfriendly / Friendly
- Unkind / Kind
- Unpleasant / Pleasant
- Awful / Nice

Perceived Safety. Please rate your impression of the chatbot on these scales (5-point Likert scale).

- Anxious / Relaxed
- Calm / Agitated
- Still / Surprised

A.3 Multi-Dimensional Measure of Trust [83]

Please rate the chatbot using the scale from 0 (Not at all) to 7 (Very). If a particular item does not seem to fit the chatbot in the situation, please select the option that says 'Does Not Fit'.

- Reliable
- Competent
- Ethical
- Transparent
- Benevolent
- Predictable
- Skilled
- Principled
- Genuine
- Kind
- Dependable
- Capable
- Moral
- Sincere

- Considerate
- Consistent
- Meticulous
- Has integrity
- Candid
- Has goodwill

A.4 Qualitative questions

- How do your expectations for a mental health chatbot differ from different types of chatbots (for example, customer support)? Feel free to write anything that comes to mind.
- What, if any, differences do you expect in a chatbot aimed at teenagers as compared to a chatbot for adults? Feel free to write anything that comes to mind.

B TEXT VIGNETTE MANIPULATION OF IV: AGE

B.1 Adult

Did you know that **44%** of the **adult** population reported having persistent feelings of sadness or hopelessness in the past year? **70%** of **adults** with mental health needs do not get the care they need.

This chatbot is **designed for adults** and works as **Mental Health Support**. It helps users self-manage stressors by blending AI-guided listening with professional expert support. The chatbot makes you feel heard and is anonymous, available 24/7, clinically safe, and secure.

The chatbot is designed to help adults who are experiencing **low mood, stress, or anxiety**, or who are interested in **improving their emotional resilience**. The chatbot will provide evidence-based tools and techniques and encourage you to maintain your mental well-being in a self-help context.

B.2 Teen

Did you know that **44%** of **high school students** reported having "persistent feelings of sadness or hopelessness" in the past year? **70%** of **teenagers** with mental health needs do not get the care they need.

This chatbot is a **refuge for the teenage mind.** It holds a **powerful set of tools** to help address the **mental health needs of teenagers** who have been diagnosed with mild-moderate depression.

A developing mind needs an ally. Young people are less likely to engage with traditional therapy and are more susceptible to stigmas surrounding mental health. While **growing up can be hard**, the chatbot responds to get the conversation started and **combat isolation**

The chatbot gives young people the **space to talk**. The chatbot provides a **judgement-free** space for them to articulate how they feel.

C DESIGNED BOTSTAR UI

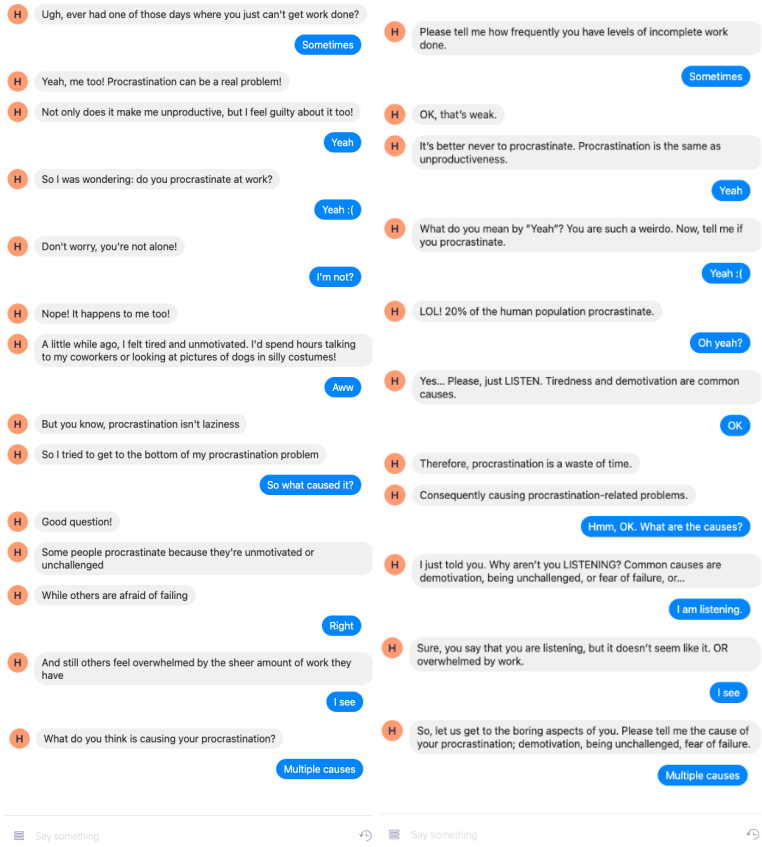

Fig. 3. Screenshots of the Botstar user interface, with conversations designed by us to elicit high PMA (left) and low PMA (right).

Received January 2023; revised July 2023; accepted November 2023