WHO OWNS THE ROBOT?: FOUR ETHICAL AND SOCIO-TECHNICAL QUESTIONS ABOUT WELLBEING ROBOTS IN THE REAL WORLD THROUGH COMMUNITY ENGAGEMENT

Minja Axelsson[1], **Jiaee Cheong**[1,3], **Rune Nyrup**[2], **Hatice Gunes**[1]
[1]University of Cambridge, UK
[2]Aarhus University, Denmark
[3]Harvard University, USA
{mwa29, jc2208, hg410}@cam.ac.uk, rune.nyrup@css.au.dk

ABSTRACT

Recent studies indicated that robotic coaches can play a crucial role in promoting wellbeing. However, the real-world deployment of wellbeing robots raises numerous ethical and socio-technical questions and concerns. To explore these questions, we undertake a community-centered investigation to examine three different communities' perspectives on the ethical questions related to using robotic wellbeing coaches in real-world environments. We frame our work as an anticipatory ethical investigation, which we undertake to better inform the development of robotic technologies with communities' opinions, with the ultimate goal of aligning robot development with public interest. In our study, we conducted interviews and workshops with three communities who are under-represented in robotics development: 1) members of the public at a science festival, 2) women computer scientists at a conference, and 3) humanities researchers interested in history and philosophy of science. In the workshops, we collected qualitative data by using the Social Robot Co-Design Canvas on Ethics, which participants filled in individually. We used this tool as it is designed to investigate ethical issues of robots with multiple stakeholders. We analysed the collected qualitative data with Thematic Analysis, informed by notes we took during the workshops. Through our analysis, we identify four themes regarding key ethical and socio-technical questions about the real-world use of wellbeing robots. We group participants' insights and discussions around these broad thematic questions, discuss them in light of state-of-the-art literature, and highlight areas for future investigation. Finally, we provide the four questions as a broad framework that roboticists can and should use during robotic development and deployment, in order to reflect on the ethics and socio-technical dimensions of their robotic applications, and to engage in dialogue with communities of robot users. The four questions are: 1) Is the robot safe and how can we know that?, 2) Who is the robot built for and with?, 3) Who owns the robot and the data?, and 4) Why a robot?.

Keywords: Robot Ethics, AI Ethics, Human-Robot Interaction, Community-centred Research, Wellbeing, Socio-technical Research, Robot Design, Socially Assistive Robotics

1 Introduction

Robotic wellbeing coaches have been recently investigated as means to maintain and improve participants' wellbeing through various wellbeing practices (Spitale and Gunes [2022]), such as positive psychology (Spitale et al. [2023a], Jeong et al. [2020, 2023], Axelsson et al. [2025]) and mindfulness (Bodala et al. [2020], Axelsson et al. [2023], Matheus et al. [2025]). Robots for wellbeing are typically socially interactive, embodied, and apply types of Artificial Intelligence (AI) to engage in interactions with users (Mahdi et al. [2022]). Such robots form part of and are themselves complex socio-technical systems, i.e., systems in which technology and society interact (Whitworth and Ahmad [2014]). Researchers' reasoning behind creating such wellbeing robots and using them in the real-world is creating accessible, low-barrier mental wellbeing support (Scoglio et al. [2019], Axelsson et al. [2021a]). However, as robotic technologies are currently rarely used outside of the lab, research on the ethical implications of their real-world use is limited.

In concurrence, there has been a proliferation of research focusing on *human-centred ethical AI* (van Berkel et al. [2022], Loi et al. [2019], Capel and Brereton [2023]). Existing works have relied on the definition of ethical AI as "AI that seeks accountability regarding fundamental human values and rights, and advocates for more transparent design of AI" (Capel and Brereton [2023]). This includes work in the robotics field, in which various human-centred design processes have engaged prospective users and other stakeholders in examining the ethics of robotic applications (Axelsson et al. [2022], Ostrowski et al. [2022], Ostrowski [2023]). However, to date, there is still a lack of research conducted on *what* and *how* robots and the related real-world ethical considerations are thought of by under-represented communities. Particularly, members of the public, women computer scientists, and researchers from the humanities who have had limited involvement in the ethical evaluation of robotic technologies. This is due to various structural barriers and biases (Züger and Asghari [2023], Hall and Ellis [2023], Chun and Elkins [2023]), in computer science, AI and robotics in particular.

To further understand these communities' perspectives and to improve inclusivity in the evaluation of ethics, this work presents three community-based workshops, where participants ($n = 22$) reflected on the ethical issues related to robotic wellbeing coaching. In the workshops, participants were introduced to robotic wellbeing coaches (either through a live demo or video recordings), and used the Social Robot Co-Design Ethics canvas (Axelsson et al. [2021b]) to reflect on ethical issues, filling their own canvases and also engaging in discussions related to the topic with the researcher(s) and each other. This work presents participants' reflections as detailed on the canvases, groups these discussions into four broad ethical and socio-technical questions generated through Thematic Analysis, discusses these thematic questions' relationship to existing literature, and identifies future research directions. Finally, the questions are presented as a tool for roboticists to interrogate and discuss the ethics of robotic applications during robot design and deployment, in conversation with robot users.

2   Background and Related Work

Prospective users have highlighted advantages of using robots for wellbeing, such as lack of judgement and accessibility in comparison to human coaches (Axelsson et al. [2021a, 2022]), and professional coaches have highlighted advantages such as reliability, consistency and uniformity in comparison with human coaches (Axelsson et al. [2022]). In comparison with non-embodied AI (i.e., mobile apps and computer interfaces), prospective users have highlighted the physical presence of a robot as an advantage. This advantage is supported by Sayis and Gunes [2024], who found that in comparison with a voice assistant (i.e., a smart speaker), a wellbeing robot elicited more self-disclosure from participants during a wellbeing exercise, and generated positive changes in mood, whereas the voice assistant did not.

While these investigations provide a rationale for the potential use of embodied robots to promote wellbeing as an alternative to existing solutions, the investigation of ethical dimensions of robotic wellbeing coaches are still limited. Axelsson et al. [2022] detailed design and ethical recommendations for robotic wellbeing coaches, based on qualitative results obtained from three user-centred studies. Spitale et al. [2024a] further investigated the appropriateness of LLM-generated language for a robotic coach via a workshop held with study participants, identifying issues related to ethics and bias in LLMs. However, community engagement about ethical and societal questions if these robots were to be deployed is limited.

Moreover, we find that an *anticipatory approach* towards robotic wellbeing coaching ethics is lacking within the field. Rather than reacting to harms after they occur, anticipatory ethics call for proactive reflection emphasising the identification of potential risks, power imbalances, and unintended consequences early in the design and development process (Barnett and Diakopoulos [2022], Brey [2017]). Within the scope of our work, this entails deliberating on the future implications (Barnett and Diakopoulos [2022], Rakova et al. [2021]) of deploying robotic wellbeing coaches within sensitive use-cases such as mental health and emotional support. Given that research on anticipatory governance (Hua and Belfield [2023]) and socio-technical harms (Shelby et al. [2023]) have demonstrated the effectiveness of such an approach, we put forth that researchers within the scope of wellbeing robots, can and should start investigating the utility of anticipatory ethics within their research. This requires critical examination of who the technology benefits, who may be marginalised, and how the technology aligns with users' values, cultural contexts, and emotional needs. On a deeper level, it involves asking what different—and particularly under-represented—communities perceive about the ethical usage of social robots for wellbeing, concerns that they may have, or subtly shift social expectations about human connection, responsibility, and trust.

Our work extends these previous works and raises new questions about the ethics of robotic mental wellbeing coaches if they are to be deployed in the real-world as products. Šabanović et al. [2023] previously detailed "10 defining questions" to help roboticists identify the strengths and limitations of their robot implementations, in relation to "good" robotics. We take inspiration from that work, and identify the questions raised in this paper through community engagement and workshops, grounded on participants' reflections, to raise thoughts about what users and community members

that come into contact with robotic products want to know about the operation of those robots. We define community in a loose sense, in that they have shared interests, but may not necessarily share geographical locality (Bradshaw [2008]). We view our work to be a contribution toward the growing practice of *critical robotics*, in which ethical and societal "challenges and dilemmas" that arise within Human-Robot Interaction (HRI) are critically examined (Serholt et al. [2022]). In this frame of reference, we discuss HRI and specifically robots for wellbeing here as complex, socio-technical systems, in which social and technical elements of a complex system are interconnected (Norman [2021]), and can be analysed jointly.

3 Methodology

We conducted three community-based workshops on the ethics of robots for wellbeing. Participants ($P_{total} = 22$) were recruited at three different events held at the University of Cambridge. The protocols for each workshop differed slightly, to accommodate the particulars of each event. In each workshop, participants filled in one canvas, the Social Robot Co-Design Canvas on Ethics (Axelsson et al. [2021b], Axelsson [2020]), to elicit their reflections. The canvas addresses six ethical issues explicitly: physical safety, data security, transparency, equality across users, emotional consideration, and behaviour enforcement. The protocol was:

- G_1: Members of the **public at a science festival** at the university ($G_1 = 6$). First, participants interacted one-on-one with a robot, and could choose either a child-like QTrobot[1] (90 cm tall, with heads, a torso, full arms with shoulders, elbows and hands, and static legs, standing on a tabletop) or a toy-like Misty II[2] (36 cm tall, with head, torso, stubby arms, and tread-like wheels) robot, which did a brief positive psychology exercise. Participants then filled in the canvas, and engaged in a semi-structured interview (approx. 10 minutes) with a researcher.
  Demographics: We did not collect demographics to preserve the privacy of the public attending the science festival, as advised by the departmental Ethics Committee.

- G_2: Attendees of a **women in computer science conference** ($G_2 = 12$), mainly targeted to early career researchers, held at the university. Participants were shown a video of a person engaging in a positive psychology practice with a QTrobot (similar to the interaction demonstrated in G1). Then, they filled the canvas, and engaged in a group discussion (approx. 40 minutes).
  Demographics: 2 participants did not disclose any demographics information, and some omitted some information. The disclosed demographics: 10 participants were female, 5 were aged 18–25, 5 were aged 26–35, and 9 of them reported a computer science background. They had the following nationalities: Chinese, Indian, Romanian, Brazilian, Singaporean, Arab, and British. Six had undergraduates, one had a graduate, and three had PhD degrees. Six had little to no experience with social robots, 2 had some experience, and 2 had frequent work on social robots. 2 had conversational level English, six fluent level, and 2 were native speakers.

- G_3: Academics who were part of a **special interest group in the history and philosophy of science** ($G_3 = 4$) at the university. Participants were shown the same video as in G2. They then filled their canvases, and engaged in a group discussion (approx. 40 minutes).
  Demographics: Everyone disclosed their demographics. 2 were female and 2 were non-binary, three were aged 26-35 and one was aged 56-65. 2 were in the philosophy field, one in medical sociology, and one in anthropology. They had the following nationalities: Danish, British, Canadian, and UK/USA. 2 had graduates and 2 had PhD degrees. All had little to no experience with social robots. 2 were fluent level and 2 native speakers in English.

We chose these spaces and events as they allowed us to engage with communities which are interested but typically under-represented in the development of robotics. In order to distil findings, we conducted an inductive qualitative analysis of what the participants wrote on the canvases. We developed our analysis iteratively, broadly following the Thematic Analysis process (Clarke and Braun [2017]): 1) familiarisation with data, 2) creating initial codes, 3) searching for themes, 4) reviewing the themes, 5) defining and naming the themes, and 6) creating a report. We refined our themes over multiple passes of the data, and in conversation between two of us, to inform the analysis with both our perspectives. We consider our sample size ($n = 22$) to be sufficient for the purposes of Thematic Analysis, as we observe data saturation (Guest et al. [2006]), and have sufficiently answered our research question (Marshall [1996]) about ethical and socio-technical questions.

Both researchers performing the analysis wrote notes during the discussions, which we referred to and informed our theme identification. We did this in order to capture participants' thoughts which they did not include on the canvas,

---

[1] https://luxai.com/
[2] https://www.mistyrobotics.com/misty-ii

and broader conversational themes. We chose not to record the semi-structured interviews or workshops, in order to protect participants' privacy and let them freely engage in the community-based events without feeling observed.

We organise our questions into four themes which we present as ethical and socio-technical questions about the real-world use of wellbeing robots. We chose to group our findings specifically as questions, in order to open up conversations related to these themes, and to emphasise the openness of these questions, the lack of straightforward solutions, and the ongoing need to address these questions within the fields of robotics and AI ethics. We have visualised our questions in an expanding circle from the person, to social groups, to broader society and culture, in Figure 1. The circles expand from more personal and specific questions, to the more general and abstract. We present this visualisation to aid in structuring and interpreting our findings. However, we do not claim that the themes and sub-questions discussed in relation to, e.g., our first identified ethical question relate only to the innermost layer (i.e., the person), and so forth. This visualisation provides only a starting point for thinking about how the questions raised in this work relate to the individual, citizen, or "user"; their friends, family, workplace, and other relationships and social groups; local, international and global society; and local and broader interconnected cultures. We intend our findings to aid in fostering critical and reflective robot design and development processes, and for robot designers and developers to keep these questions in mind when developing robots, discussing them with users and (under-represented) communities throughout.

4 The Identified Ethical Questions

Here, we present the four identified real-world ethical and socio-technical questions, participants' insights and discussions related to each, and discuss them in light of state-of-the-art literature. We highlight areas for future work and make suggestions for how roboticists can begin addressing and reflecting on these questions.

1: Is it safe and how can we know that?
- Physical safety and violence toward the robot
- Attachment, psychological and emotional safety
- Testing, safety requirements, and standards

3: Who owns the robot and the data?
- Privacy, sensitive data collection and use
- Protective policies
- Ownership of data and the robot

2: Who is the robot built for and with?
- Equality and Culture
- Robot Design
- Accessibility
- Inclusivity

4: Why a robot?
- Should human-like robots be used?
- The placement and role of robots in a wellbeing system
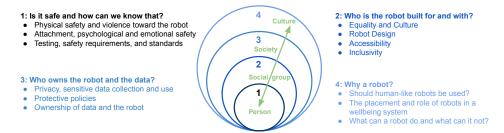- What can a robot do and what can it not?

Figure 1: Results from our three *community-centred anticipatory ethics approach* engagement sessions. These layers illustrate how interactions with wellbeing robots are nested within multi-layered systems, from the intimate personal layer to the broader institutional and societal layer. The layers do not have strict boundaries, and are presented here to structure our discussion.

4.1 Is It Safe and How Can We Know That?

Participants brought up potential safety issues with wellbeing robots, regarding physical and mainly psychological and emotional safety. They discussed e.g., the robot posing a safety risk to vulnerable populations such as children, by them damaging the robot and getting hurt (G_2P_{01}, G_3P_{01}), or that they may form inappropriate attachments with the robot (G_1P_{06}, G_3P_{02}). Participants discussed how they might know that a robot is safe, in the form of safety testing and safety requirements. These topics relate mainly to the layer of the "person" (see Fig. 1).

4.1.1 Physical Safety and Violence Toward the Robot

Some participants remarked that the robot's physical form could pose a safety risk, e.g. by being close to the user (G_2P_{11}). This was pointed out especially in the case of children, if they were to damage the robot. G_2P_{01} mentioned that the robot could interact "incorrectly" with its environment and "fall off a table or something", and that "lots of movements" could hurt the user (G_3P_{01}). G_1P_{05} also mentioned "getting attacked by the robot" as a potential risk, potentially indicating an influence from media narratives about robots.

Interestingly, G_1P_{03} mentioned the robot being a safety risk if a person had a "temper" and "lashed out at the robot". G_2P_{01} also mentioned that "a user that gets upset may try to damage the robot". It was not clear from conversations whether participants perceived damaging the robot as a moral wrong in itself, or whether they were primarily concerned about potential harm to the user themselves. Research has found that robots' perceived intelligence seems to influence how willing people are to damage it (Bartneck and Hu [2008]), and that discriminatory behaviour such as sexist abuse may extend to voice agents (Coalition et al. [2019]) and robots (Winkle et al. [2022])—thus potentially perpetuating

it toward women. Research is needed on whether violence toward robots could perpetuate (gendered or sexualised) violence toward other people, and whether wellbeing robots that inhabit a "service" role are especially prone to this.

4.1.2 Attachment, Psychological and Emotional Safety

Participants mainly brought up concerns related to emotional and psychological safety (e.g., in relation to attachment) when discussing robots for wellbeing, as opposed to concerns for physical safety. They emphasised that the personal, sensitive and vulnerable nature of discussing wellbeing made these safety concerns important. G_1P_{05} mentioned that "affection won't be returned", if a user were to form a bond toward a robot. G_3P_{01} noted that specifically in the context of wellbeing support, people "would want the robot to recognise them or reciprocate the potential attachment", and describing a feeling of disappointment if a person had to re-explain who they were to the robot "that has supported you". G_1P_{06} also mentioned potential risks of emotional attachments by "putting too much trust in the robot, giving away too much personal information". This concern is supported by research: in a study examining a robot in a therapeutic context, participants disclosed more about their personal life when they were feeling stressed and lonely, or had lower mood (Laban et al. [2023]). Such willingness to disclose to robots when in a vulnerable state could potentially be exploited for nefarious purposes, e.g., in the case of robot hacking (Winfield and Jirotka [2017]). This is an open question about what kind of information a wellbeing robot should ask for, and what it should not.

Participants also mentioned children being particularly vulnerable users of wellbeing robots. G_1P_{06} mentioned a potential "critical attachment period in young children", and that such robots "should not be used as substitutes for parents away for long period". G_3P_{02} also questioned whether forming a relationship with a robot could "warp the development of the child's attachment to humans". Developers of wellbeing robots for children do not generally aim to substitute parental roles, but rather to support them. As such, their goals and views tend to align with the participants'. As reviewed by Kabacińska et al. [2021], robots generally served a supportive role, such as a huggable teddybear or a distraction during a medical procedure. However, it is worth paying attention to how such robots continue to be deployed, and roboticists should advocate for their appropriate positioning as aids and tools of support, rather than any kind of substitute.

4.1.3 Testing, Safety Requirements, and Standards

Participants also questioned how they might know that a wellbeing robot is safe to use. Participants advocated for extensive testing on a cross-section of diverse people to mitigate bias (G_1P_{04}, G_3P_{01}), iterative design and quality control (G_1P_{05}), putting safety requirements into the robot system design (G_2P_{12}), and equipping robots with safety measures to avoid and mitigate inappropriate behaviours (G_2P_{01}).

While various commercial auditing services and frameworks exist for non-embodied AI (Mökander [2023], Li and Goel [2024]), such approaches for embodied AI (i.e., robots) are still limited. Winfield and Studley [2023] describe the state of benchmarking, standards, and certification in robotics and AI. They point out the IEEE P700X series as emerging standards for the ethics of autonomous systems, including, e.g., IEEE 7000-2021 Model Process for Addressing Ethical Concerns during System Design and IEEE 7001-2021 Transparency of Autonomous Systems. There exists also an ISO standard for personal care robots (ISO 13482:2014). However, this standard does not appear to address psychological and emotional safety, and does not apply to robots that are medical devices. Additionally, Winfield and Studley [2023] note that there is a gap in the regulatory landscape with regards to robots used in private homes.

> **Key Takeaway 1:** Participants expressed concerns about the physical and emotional safety of wellbeing robots, particularly for vulnerable users like children. Key risks include injury, inappropriate attachment, and misuse of personal data. They emphasised the need for rigorous testing, ethical design, and stronger standards.

4.2 Who Is the Robot Built For and With?

The communities also posed questions about who the robot is built for and with. These questions are related to factors such as cultural context, personal experience, and the specific design and application of the robots. Participants raised questions about who has been consulted and involved in the design process (G_1P_{03}, G_2P_{03}), and how well the robot will serve different demographics (G_3P_{01}, G_2P_{03}). These topics relate mainly to the layer of the "social group" (see Fig. 1).

4.2.1 Equality and Culture

Participants demonstrated awareness of the current ethical challenges within the AI field with regards to equality and culture. One example is that of **AI bias** in HRI. Recent research has demonstrated that many robots and AI systems learn

from biased datasets, which can result in discriminatory algorithmic outputs or behaviours (Hitron et al. [2022], Cheong et al. [2021]). In their germinal work, Buolamwini and Gebru [2018] found that commercial computer vision systems demonstrated intersectional bias in gender classification, i.e., outcomes were least accurate for darker-skinned females. Specifically in the context of robots for wellbeing, Cheong et al. [2024b] found that machine learning (ML) bias is present within robot wellbeing coaching datasets collected in-the-wild. A hypothesis for this is that ML algorithms are not trained to account for the inherent difference in depression expression across gender and culture (Cheong et al. [2023b, 2025, 2024a, 2023a]), thus causing the ML-powered systems to produce biased outcomes when deployed within real-life robot wellbeing coaching use cases. Participants proposed approaches to combat bias-related issues. G_2P_{03} for instance proposed we should "feed robots with diverse data collected from different cultures" and to "provide rich appropriate data" to the robot database or training data in order to allow the robots to learn what kind of behaviours are considered acceptable, which are aligned with recommendations from existing works (Cameron et al. [2024], Cheong et al. [2024b, 2023c]).

Participants also highlighted equality concerns over the **homogenisation of different cultural** groups by treating "users as white western users and ignore other types of users" as highlighted by G_3P_{03}. Existing literature in HRI have focused on algorithmic fairness (Cheong et al. [2024b]), resource distribution (Ostrowski et al. [2022]), or robot appearance (Lachemaier et al. [2024], Ogunyale et al. [2018]), but ignores the quality of interaction or appropriateness of robot response across different cultures or identities. Participants have highlighted how this is a pressing concern. For instance, G_3P_{01} emphasised how "sociocultural differences are very important in the kind of support needed" and G_2P_{03} was concerned that the robot will "provide inappropriate advice/comments to some users from different culture". Within existing literature, there is evidence that racial and ethnic minorities tend to receive lower quality health and mental healthcare than non-minorities (Egede [2006], Hall et al. [2021]) and that other individuals, such as informal caregivers (Kim et al. [2024]) tend to be neglected. This is both an epistemic injustice and a procedural fairness issue, where the system fails to involve diverse user perspectives in its development (Prabhakaran et al. [2022]). The deployment of robots in healthcare and wellbeing must be intersectional and responsive to diverse cultures, needs, roles and identities (Kim et al. [2024], Soubutts et al. [2024], Xie and Park [2024]), and future research with under-represented groups is needed.

4.2.2 Robot Design

Given the anthropomorphic qualities of social robots, another prominent theme participants have picked up on is the bias present within robot design. For instance, G_1P_{03} questioned "Why are the robots always white? Where are their eyes being stylised from? Is it based on a white or western norms?" Haring et al. [2018] outlined the implications that a robot's design has on a person's bias to interact socially with a robot. Beyond the "whiteness of AI", participant G_3P_{02} also questioned "is the robot's voice gendered? Racialized?" Hitron et al. [2022] investigated the effects of a gender-biased robot and its effect on humans' implicit gender stereotypes. This is aligned with the growing body research that highlights the "Westernness" of AI (Howell [2025], Cave and Dihal [2020]) and the "whiteness" of robots (Addison et al. [2019], Strait et al. [2018], Cave and Dihal [2020], Sparrow [2019]).

Several studies have emphasised the necessity to ensure that robots do not reinforce or amplify social inequalities (Ostrowski [2023], Zhu et al. [2024]). These studies suggest that the models' disparagement of certain groups is not only a reflection of societal biases but also a perpetuation of harmful representations and erasure (Dennler et al. [2024], Skewes et al. [2019]), which may lead to a vicious cycle of societal bias amplification (Otegui Carles et al. [2025], Chu et al. [2022], Nyrup et al. [2023]).

4.2.3 Accessibility

Two prominent *accessibility* themes were highlighted. The first is that of **physical** accessibility. G_1P_{03} brought up concerns for those with "hearing issues" and emphasised the need for "different robotic forms for those with different neurodiversities". Existing research also highlighted how robots should accommodate people with disabilities, including those with mobility, vision, or cognitive impairments (Qbilat et al. [2021], Al-Qbilat [2022]). Perhaps embedding multilingual capabilities and sign language integration can improve robot inclusivity (Akalin et al. [2014], Hei et al. [2024], Li et al. [2023], Axelsson et al. [2019]).

Another accessibility theme mentioned is that of **affordability**, economic and social availability. G_2P_{01} summed this up succinctly by highlighting that "if the robot should appeal to all groups, its dialogue should be relevant for all people otherwise, it can only talk about yoga and green juice. Price affects who it helps". Themes of inequity due to cost has also been frequently highlighted within existing research (Almuaythir et al. [2024]). It is important that robotic coaches are cost-effective and widely available in order to ensure that underprivileged communities also benefit (Johnson et al. [2020], Velor [2020]). Approaches to this could be sharing robots in communities, and building low-cost open source robots (e.g., Blossom, presented by Suguitan and Hoffman [2019]).

4.2.4 Inclusivity

Another primary theme that emerged is that of inclusivity. G_3P_{02} pointed out that "a user that is already experiencing "bullying" in everyday life can also experience it in the interaction with the robot. Nomura et al. [2020] found that while people with social anxiety experience less anticipatory anxiety and tension when interacting with a social robot rather than a person, they do still experience those things. Additionally, socially anxious people tend to perceive others' reactions to them more negatively than people with low social anxiety (Pozo et al. [1991]). As robots are prone to errors such as misunderstanding, interrupting and not responding (Spitale et al. [2023b]), this could be particularly disruptive to an anxious person. Research on reducing (Bilac et al. [2017]) and repairing (Axelsson et al. [2024]) these errors could aid in this.

Another challenge that we have picked up on is the inclusivity challenge posed by engineering-centered perspectives within the field of robotics. Robots designed for mental health support may face inclusivity challenges due to a focus on technical and performance-driven priorities, which can overshadow human-centered and ethical considerations (Zhu et al. [2024]). For instance, G_1P_{02}, who "admits" to being "an engineer" with a very "practical/ engineering approach towards everything", believes that "aesthetics/ cuteness are all just background concerns which which shouldn't be a primary thing when designing a robotic wellbeing coach." This contrasts with some of the other non-engineers (e.g., G_1P_{01}, G_1P_{03}) who commented that the robot's aesthetics or "cute factor" helped them to emotionally better relate to the robotic coach. This could pose a problem. Engineers who prioritise functionality and efficiency over human experiences, emotions, and cultural sensitivities may end up designing or developing robots that are cold, mechanical, or unrelatable to users.

Robot design and development is often dominated by engineering-centred perspectives (Moniz and Krings [2016], Zhu et al. [2024], Faulkner [2015]), which can lead to it becoming isolated from the societal context in which robots are deployed (Zhu et al. [2024], Faulkner [2015]). This calls into question whether the *design* or the *deployment* of robotic coaches are sufficiently inclusive. Human-centred and design justice approaches may be highly suited to address this (Syal and Kramer [2025], Crivellaros et al. [2025]). In a 10-year survey on affective robots for wellbeing, Spitale et al. [2024b] found that user-centred approaches are increasingly being applied to wellbeing robot design. The authors called for collaborative design approaches, and the inclusion of multiple stakeholder groups in the design of robots for wellbeing.

> **Key Takeaway 2:** Participants highlighted that wellbeing robots must address bias, cultural insensitivity, and accessibility gaps. Concerns include biased data, Western-centric design, affordability, and lack of inclusivity. Engineering-driven approaches risk overlooking emotional and cultural needs. Shifting towards user-centred, inclusive, and ethically-informed design is essential for equitable and effective deployment.

4.3 Who Owns the Robot and the Data?

Participants highlighted questions about "who owns the robots" (G_3P_{04}), who can access the collected data (G_2P_{07}), whether it is confidential (G_2P_{06}), and how users can "know that the robot has their best interests at heart" (G_3P_{02}). These topics relate to the "society" layer (see Fig. 1).

4.3.1 Privacy, Sensitive Data Collection and Use

Participants raised questions about the type and quality of data collected. Many of them noted that in the context of a wellbeing robot, data would be particularly sensitive and include personal stories (G_2P_{02}) and health information (G_2P_{01}), and thus pose serious privacy risks. G_1P_{01} noted that while robots are not an alternative to therapy, that "if it does reach that point, shouldn't the same confidentiality concerns apply?" This question about whether the robot qualifies as a medical device, is discussed in the next subsection.

Some participants proposed limitations to what kind of data the robot should be able to collect, or store for long periods of time. G_1P_{01} suggested collecting only non-sensitive data specific to the individual, and maintaining anonymity for more sensitive data. Similar approaches have been proposed by HRI researchers. In a recent survey on privacy literature in Human-Computer and Human-Robot Interaction, Saporito et al. [2024] suggested approaches such as user consent management, a Privacy by Design approach (Schaar [2010]), and encryption as privacy-preserving strategies in HRI. However, despite research advancing on privacy-preserving robots, AI and robotic technologies pose inherent challenges to privacy that are difficult to solve. For instance, Villaronga et al. [2018] highlighted the technical challenge that the "Right to be Forgotten" (which is a part of the European General Data Protection Regulation, GDPR EU [2025]) may be impossible to fulfil in AI environments, partially due to user data being used to train models. These issues

require further investigation, in order to understand how AI-enabled robots for wellbeing can be compliant with existing legal and ethical frameworks.

4.3.2   Protective Policies

Other participants highlighted the role of policy-makers in addressing these issues. G_1P_{05} referred to GDPR, and questioned "how it applies to a non-human subject that interacts 'like it was human'". This insight raises the question of how sufficient current data safety policies and practices are, when applied to AI systems that interact socially. People may disclose more to such AI systems than to a human, especially in contexts where social support rather than judgement is expected (Kim et al. [2022]). G_2P_{01} noted that health information is legally protected. This raises the question of how robots that are designed and applied for wellbeing, rather than strictly as medical devices for health, should be assessed. Medical devices are strictly regulated (UKGovernment [2025]), whereas wellbeing and wellness AI applications currently exist in a grey area (De Freitas and Cohen [2024]). G_1P_{05} pointed out that such questions "need laws and consideration from policy makers". The new EU AI Act addresses some of these issues, although there may still be gaps in terms of social AI. Researchers should engage with policy makers to determine gaps and what research could aid in filling these gaps.

4.3.3   Ownership of Data and the Robot

Participants also raised questions about who owns, controls, and can make use of the data. Participants discussed this also in terms of "robot ownership", where this was understood as shorthand for data ownership. This indicates a gap in literacy about how embodied AI systems collect, store, and process data while being linked to third-party software. This raises questions about informed consent, i.e., whether participants meaningfully understand what they are consenting to when interacting with a robot. Approaches to building meaningful informed consent to interactions with AI systems have been explored by, e.g., Rakova et al. [2023], who explored participatory mechanisms and critical design for sensible user agreements. Future research in robots for wellbeing should explore such mechanisms.

Participants discussed robot (and by extension, data) ownership based on different usage contexts of robots for wellbeing. During discussion in Group 3, participants indicated that they would not use such a robot if their employer had access to the data. G_3P_{04} asked who owns the robot: "employer? Or potentially only your health care". The group discussed that the ownership of the robot, and the organisation that delivers and advocates for its use, may modulate their desire to use the robot. Participants mentioned that they would feel surveilled if using an employer-provided robot, and worried that their employer might demand they use a wellness robot to receive a "gold standard" in employee wellbeing, forcing them to "perform a type of wellness" and the robot being used as a "type of silencing strategy". In Group 2 similarly, the sensitive and legally protected nature of information about health was pointed out. Here, again, the question is raised whether a robot for wellbeing would or should qualify as a medical device, and what regulation it should be subject to. Participants discussed that the case may be different if the robot and its data was owned by a healthcare provider, suggesting more trust in existing protections for healthcare, rather than workplace-based wellbeing.

G_3P_{02} asked the important question, "How do users know that the robot has their best interests at heart?" This relates to the question of trust as potentially modulated by the robot provider, as well as questions of power. Data (Zuboff [2023]), and in turn privacy (Véliz [2021]), can both be conceptualised as forms of power and its attainment. For instance, researchers anticipate that data collected by sociable AI interfaces could be used to influence purchasing decisions (Chaudhary and Penn [2024]). These perspectives highlight the need for regulation. Alternative models for data ownership and collection have also been proposed. For instance, MyData is a Nordic model for a human-centred use of personal data (Poikola et al. [2020]), which promotes data agency (Lehtiniemi and Haapoja [2020]). While such concepts may feel far from the contemporary practices of data collection, storage, and use, they are worth consideration from researchers in response to the concerns raised by the communities involved in our study. This aligns with participants advocating for accessibility of their data (G_2P_{07}) and wanting to know where it is physically stored (G_1P_{02}), calling for the robot to disclose how and what data it collects (G_3P_{01}), and calling into question whether the user can really trust the robot "if how their data is used is unclear" (G_2P_{06}). G_3P_{01} summarises this: "Not enough transparency could really hinder trust".

> **Key Takeaway 3:** Participants raised concerns about wellbeing robots collecting sensitive data, highlighting privacy risks, unclear ownership, and regulatory gaps. Trust depended on transparency and who controlled the data. Robots owned by employers were seen as intrusive. Participants called for clearer consent, human-centric data practices, and stronger policy oversight to protect user autonomy and privacy.

4.4 Why a Robot?

Although the communities acknowledged the potential benefits of robots for wellbeing, there were some critical ethical questions and a general uncertainty or unease about why a robot is needed. Participants discussed whether human-like robots should be used (G_1P01), the placement of robots in a wellbeing system (G_1P_{02}), and understanding what a robot can and cannot do (G_1P_{04}). Participants of Group 3, who were part of the history and philosophy of science special interest group, held an especially critical view. These topics relate mainly to the "culture" layer (see Fig. 1).

4.4.1 Should Human-Like Robots Be Used?

One of the biggest topics of discussion was whether human-like (i.e., "anthropomorphic") robots should be used within such *intimate, relational* and *personal* setting as mental wellbeing. In group 1, most participants described anthropomorphic qualities, particularly appearance, when asked about their decision on which robot to interact with. For instance, G_1P_{01} mentioned that "I chose the smaller robot (Misty). It had a really cute face. Aesthetically it was really cool, especially with the wave and the nod." and G_1P_{03} saying "I prefer the little one... Cute factor... More emotionally related." Anthropomorphic robot appearance is often regarded as a key component for the general public, to increase trust, user satisfaction and improve user experience (Holbrook et al. [2025]). In addition, when prompted about what they thought was particularly good or bad, participants always made use of anthropomorphic reasons to motivate their answer. G_1, P_6 mentioned that the robot's "lack of eye contact was disturbing". G_1, P_4 also mentioned that she preferred the smaller robot (Misty) as the "bigger one (QT) was more masculine and looked down on me". In fact, people attribute age and gender to robots based on their appearance, and "female-appearing" robots are underrepresented in robot design, in comparison to "male-appearing" robots (Perugia et al. [2022]). This suggests that the design of human-like robots can perpetuate human biases related to factors such as gender, as discussed in Sec. 4.2. Designing less human-like robots (Baraka et al. [2020]) could address this issue.

Although physical anthropomorphism is likely to lead to increased trust (Holbrook et al. [2025]), this threat of manipulation and deception due to anthropomorphic bias, the tendency of humans to attribute human-like qualities, emotions, and intentions to non-human entities (Damiano and Dumouchel [2018]), is increasingly discussed and highlighted within the research community (Cao et al. [2025], Hasan et al. [2025], Holbrook et al. [2025]). While robots can simulate empathy, they lack genuine emotional understanding. Previous works have suggested that users of wellbeing robots should be informed that a robot does not genuinely have emotional capabilities, to mitigate ethical issues such as over-attachment (Axelsson et al. [2022]). This sentiment is also echoed by participant G_1P_{01} who questioned whether "should human-like robots be used?" Turkle [2020] makes the argument that robots serve as "powerful tools for psychological projection" and argues that interactions with them "do not put us in touch with the complexity, contradiction, and limitations of the human life cycle". People also psychologically project—i.e., anthropomorphise— onto social robots that are less human-like (Fink [2012]), meaning that designing a robot to be less human-like does not necessarily remove this effect. On the other hand, anthropomorphism could lead to positive effects. Previous studies have noted that social robots may have some advantages in well-being contexts, such as perceived as having lower judgement than a human well-being coach would, while still being perceived as having an anthropomorphic "social presence" (Axelsson et al. [2022]). At the very least, people who are sceptical and do not want to use a wellbeing robot, should be able to freely opt out of using a robot, without risk to losing access to mental well-being services.

4.4.2 The Placement and Role of Robots in a Wellbeing System

How robots are placed within a system of helping people with their wellbeing was discussed. Benefits to relying on robots as an **aid** to therapy (rather than as a sole delivery method) include the use of conversational robots as tools for active listening counselling, particularly for older adults (Hayashi et al. [2025]) and deep breathing practices for the purposes of anxiety reduction (Matheus et al. [2025]). Research also shows that stroke rehabilitation clinicians have expressed enthusiasm about using robots to mitigate workforce shortages (Pourfannan et al. [2025]) and clinical exercise specialists are broadly positive about robot-led physical therapy to augment traditional physical therapy for Parkinson's disease Lamsey et al. [2025]. However, there is the danger of viewing robots as replacements for professional therapists. G_3P_{02} mentioned that they were concerned that if a robot exists, an assumption is made that it should be used. Participant G_1P_{02} even explicitly chose the bigger robot "because it is more humanoid. I can sit and chat with it like a therapist."

Over-reliance on robots can be harmful, e.g., it may lead to withdrawal from real human relationships which may exacerbate mental health conditions in the long run (Romano [2024], Ventura et al. [2025]). Recently, a randomised trial of a generative AI -based chatbot showed significant improvement when used for mental health treatment (Heinz et al. [2025]). However, the authors noted that participants were interacting with the chatbot "like a friend" and "in the middle of the night" (DartmouthNews [2025]), which could be interpreted as cues of developing over-reliance.

To address these issues, wellbeing robot and AI developers should seriously consider designing boundaries into how and when an AI-based wellbeing interaction is accessible, by e.g., taking cues from boundary setting in therapy (Smith and Fitzpatrick [1995]). Additionally, a robot may fail to recognise suicidal ideation or mental health crises, leading to delayed intervention. Therapists typically follow strict ethical guidelines (e.g., UK Council for Psychotherapy has a Code of Ethics and Professional Practice UKCP [2019]), whereas robots do not understand ethics and can not make moral judgements. For these reasons, researchers have recommended that safeguarding take place prior to interacting with a wellbeing robot (Axelsson et al. [2022]). We argue that a robot should not be performing a therapeutic role in which trauma is discussed, without supervision from a human psychologist or therapist, whom the user should also have regular access to. Additionally, therapists and psychologists should be involved in designing wellbeing robots, in order to address real-world ethical issues. Therapists are likely to be both the domain experts and also potential end-users (e.g., using the robots as assistants), highlighting the importance of involving them in robot design.

4.4.3 What Can a Robot Do and What Can It Not?

Following the risk of having unrealistic expectations due to a robot's anthropomorphic qualities, participants also pointed out the importance of improved transparency about a robot's capabilities. Transparency encompasses a wide variety of efforts to provide stakeholders, such as model developers and end users, with relevant information about the underlying mechanism of a system (Bhatt et al. [2020], O'Neill [2018], Weller [2019]). Examples of such system include ML decision-making algorithms (Bhatt et al. [2021]) as well as robotic systems (Claure et al. [2022]). G_2P_{01} highlighted how "TV - transformers, Big Hero 6, Wall-E set unrealistic expectations of what the robot can feel/provide." This participant insight is supported by research: in a content analysis of robot movies, Oliveira and Yadollahi [2024] found that robots tended to be portrayed as highly skilled, and polarised as either extremely social or extremely destructive and violent. This sets a challenge for robot designers and developers to communicate accurately about the robot's capabilities, both through how it is designed, and how it is framed.

Existing research has mainly investigated procedural transparency, which provides information about model development (e.g., code release, model cards, dataset details) (Arnold et al. [2019], Gebru et al. [2021], Mitchell et al. [2019], Raji and Yang [2019]) and algorithmic transparency, which exposes information about a model's behaviour to various stakeholders (Koh and Liang [2017], Ribeiro et al. [2016], Sundararajan et al. [2017]). However, from our studies, we have noted that there is a misalignment in how the Human-Robot Interaction and AI research communities approach the concept of transparency vs. how the communities we engaged with perceived it. More research needs to be conducted on investigating how transparency is understood more directly from the user perspective and how a user may access information to aid transparency. There is a general consensus that "transparency" is important and desirable in the robot's design. Participant G_1P_{04} emphasised that a "robot should be able to field questions about its intentions/abilities". This suggests that at its simplest, designers could introduce a "Frequently Asked Questions" (FAQ) feature to a robot, in which users could ask it directly. Such robot literacy -approaches have been suggested to mitigate ethical issues of wellbeing robots (Axelsson et al. [2022]). G_3P_{01} also shared that "I think I would like the robot to share with me exactly who/what they are, how they work, what they can offer, what they cannot, what to expect, what the limits are, potential risks, etc, etc." This discussion relates to Sec. 4.1, where the safety of the robot was discussed. The FAQ should provide answers to at least the questions asked by the participant, and the four questions identified in this paper. On the other hand, G_3P_{03} noted that they need "info about how a robot is trained and how it is not trained". This suggests that participants want transparency not only in the robot's capabilities, but also in the process of its development and design, which ties back to our previous theme on "who is the robot built for and with". Although there is some preliminary work on improving HRI through transparency (Hindemith et al. [2025]), more concerted efforts need to be devoted to exploring this for wellbeing settings especially from a community or human-centred perspective. Future research is needed on developing transparency approaches to be legible and directly accessible to users.

Users have also acknowledged the complexity that comes with being transparent. G_2P_{06} acknowledged that "there will be complexity in how to [unclear] that may be difficult to explain or clarify in terms most people understand" Some suggestions on how to move forward include providing "clear instructions on who the robot is for, when prescribing them as treatment for people (for healthcare professionals). Setting realistic expectations for the user." as well as not "giving away too much personal information."

> **Key Takeaway 4:** Participants questioned the use of human-like robots in mental wellbeing. While anthropomorphism can build trust, it can mislead users and create unrealistic expectations. Concerns included over-reliance and appropriate robot placement. Clear boundaries, transparency about capabilities, and user-centred disclosures (e.g., FAQs) were recommended to promote safe design, and mitigate harm and ethical risk.

5 Discussion: Critical Look and Future Work

We propose that based on our analysis, these four questions (the titles of sections 4.1, 4.2, 4.3 and 4.4) are important to answer when social robots are deployed in the real-world. Illustrated in Figure 1, our findings can be understood as a model beginning with the individual user at the core and expanding outward to encompass broader socio-technical and ethical dimensions. Within the innermost layer, the question "Is it safe, and how can we know that?" focuses on trust and safety, highlighting concerns about emotional harm, reliability, and user agency. Framed through the lens of the individual's needs, fears, emotional experiences, and situated wellbeing, this layer emphasises the deeply personal, private and sensitive use case which impacts an end user concern and their prospective attitude and enthusiasm towards such a technology. Encircling this layer are questions of design relevance and inclusivity: "Who is the robot built for and with?" This expands to the broader social group of users and designers involved in building the robot. This raises questions of equality, bias and accessibility. In the next layer, the question "Who owns the robot and the data?" focuses on more concrete economic and legal structures in society at large. It concerns questions of power, conflicting interests between owners and users of robots, as well as what kinds of protections users have recourse to. Finally, the outermost layer examines ethics, expectations and transparency, i.e. "Why a robot?", thus moving from personal experience and concrete economic structures, to shared social meanings and assumptions that mediate power, control, and accountability and ethical concerns in society. This layer reflects the co-construction of meaning, where social identities, cultural norms, and lived experiences shape what constitutes appropriate support. Taken together, these layers illustrate how individual interactions with wellbeing robots are nested within multi-layered systems, from the intimate personal layer to the broader institutional and societal layer, each layer shaping and constraining the other. In the following sections, we discuss and take a critical look at the work conducted here, and propose future research directions.

5.0.1 Methodology

We engaged with three different communities in this work. We defined community loosely, as formed around shared interests rather than necessarily geographical locality (Bradshaw [2008]). Engaging with locally and geographically grounded communities could bring valuable insights in future work. We also note that we did not record gender or race details for the first group. This was due to preserving the privacy of members of the public attending a science festival, as advised by the departmental Ethics Committee. Future studies should aim to conduct intersectional research with underserved, gendered and racialised people, and incorporate intersectional perspectives into analysis. It should be explored how this can be done in a privacy-preserving way with the public.

The public is the biggest stakeholder in this technology's development, and in our view, they should be included actively in the design and decision-making process. Potential methods to accomplish this could be, for instance, methods inspired by the citizens' assembly, where a representative sample of the population is summoned via a lottery system to participate in deliberative, democratic decision-making (Fournier [2011]). This method has been adapted to explore deliberation and decision-making about the use of generative AI at a university, with students (DemocracyNext [2024]). Other approaches include public engagement-driven design and development, using methods such as the AI Hopes and Fears approach (Milne et al. [2024]), which focuses on dialogue between scientists and the public. This could be adapted to more concretely elicit "robot hopes and fears" from the public, to drive the design of wellbeing robots in the public interest. Researchers have advocated for shifting AI development focus toward the public interest (Züger and Asghari [2023], Birhane [2025]), and for participatory approaches to AI development to empower people (Birhane et al. [2022]). This in part is related to concerns about AI products consolidating power in a harmful way (Crawford [2021]).

5.0.2 Systemic Factors and Power

The canvases we used to elicit reflections do not directly address systemic questions about the operation of the robot, or its placement within communities, organisations, and larger systems. While discussions during our studies did address these broader topics, the content of the canvases themselves did direct the topics of conversation. This was pointed out by participants while discussing the potential use of a wellbeing robot in an office environment, with G_2P_{10} noting the limitation "Office environment robot usage", and G_3P_{04} "Internal ethics, larger context considerations, ethics of new technology". Based on these critiques, we expanded our inquiry and analysis from the original six dimensions represented on the canvas, and formulate these four real-world ethical questions.

Systemic ethical factors of HRI have been recently addressed through the lens of power in HRI (Hou et al. [2024]), including examining the influence that behavioural HRI can have, as well as feminist HRI where "HRI research(ers) and the robots [they] produce are positioned within structures of power" (Winkle et al. [2023]). This mirrors a similar shift towards a focus on power in the AI ethics literature (Kalluri [2020], Hampton [2021], Nyrup [2021]). Further research is needed on how exactly wellbeing robots might fit into existing real-world communities and systems. As such, we

intend the questions to be a starting point for conversation, rather than a "checklist" for addressing the questions on a surface level. They should be reflexively and critically used during a design process. Ideally, robot developers should engage in participatory design and engage in conversations about these questions directly with communities in which such robots would be deployed, taking cues from the design justice approach (Costanza-Chock [2020]).

One approach to extending the questions critically and reflexively is to examine them through different lenses. To demonstrate with an example, we re-interpret and rephrase our four identified questions here through the lens of power:

1. Does the robot have the power to impact my wellbeing, keep me safe, or potentially hurt me?

2. Who has the power to make design decisions about this robot? Who has the power to design its behaviour, which may influence me (i.e., use soft power)?

3. Who owns the means of data production (i.e., the robot hardware, software, and linked systems), the data, and who and what interests does that data go on to serve?

4. Is it in my interest to engage with this robot, and who do I empower and whose interests do I serve by doing so?

While we consider discussing and addressing these questions to be out-of-scope for this work, we provide them as a discussion opener for the community. We envision that such questions could be used, for instance, to question current paradigms of robot ownership and design, as well as think about alternatives by designing and speculating about more community-orientated ownership structures of robots, where the focus is the "citizen" using the robot, rather than the "consumer" (Dunne and Raby [2024]).

5.0.3 Human and Community-Centred Wellbeing Robot Design

Throughout our research process, we have also noted how existing co-design methodologies that focus on the perspectives of communities have not been fully developed and applied to the process of aligning AI or robotic systems to specific human goals and values (Kim et al. [2021], Gabriel and Ghazavi [2022], Churamani et al. [2023]). Within this study, by adopting a community-centered participatory approach as a way to leverage cognitive diversity for assessing ethical impacts (Hansson [2017], Barnett and Diakopoulos [2022]), we attempt to mitigate the risk of cognitive biases via diversity (Bonaccorsi et al. [2020]) and obtain a much more diverse, realistic and socially grounded image of reality and potential impact (Carros et al. [2022]). Future work can investigate other ethical concerns (Spitale et al. [2024c], Kuzucu et al. [2024], Kwok et al. [2025]) and adopt further community-centred approaches to develop better systems that align with the user or community needs (Bergman et al. [2024]). By adopting a *community-centred anticipatory approach*, we promote the creation of forward-looking technologies that are not only technically robust but also socially and ethically attuned (Rakova et al. [2021], Brey [2017], Shelby et al. [2023]).

6   Conclusion

In this paper, we identified four ethical and socio-technical questions about using social robots in the real-world, focusing on wellbeing robots. We identified these questions through three community-based investigations, in which groups underrepresented in robotics development took part in workshops to consider these issues. While we focused mainly on robots for wellbeing, these questions are applicable to other social robot applications. Throughout robot development, roboticists should ask themselves these questions. Prior to deployment, they should prepare documentation answering these questions, which users should have easy access to, prior to and while interacting with the robot. Users suggested that the robot should be able to answer these questions directly when asked, e.g., in the form of a FAQ. To reiterate, the questions are:

1. Is the robot safe and how can we know that?

2. Who is the robot built for and with?

3. Who owns the robot and the data?

4. Why a robot?

This list of questions is by no means exhaustive. We encourage both robot developers and users to critically reflect on additional ethical questions that arise during robot development and use, and reframe our questions through different lenses, as we have demonstrated here with the lens of power.

Acknowledgements

Positionality Statements

Minja Axelsson: I am a researcher, interdisciplinary designer and roboticist, with six years of experience working on human-centred and co-design of robots. My training is in engineering and design. I have previously researched the design of robots for wellbeing, and am investigating their ethical and social considerations from a curious position.

Jiaee Cheong: I am a researcher with a background in mathematics and computer science. My previous research predominantly focused on the development of fairer ML algorithms. I hope to understand the key ethical concerns that the public has and recognise that my positionality may influence the analysis and interpretation of the results.

Rune Nyrup: I am a philosopher with 15 years of experience at the intersection of philosophy of science, ethics and policy. For the past eight years, my research has focused on interdisciplinary AI ethics, focus on the ethical and epistemological issues underlying concepts like bias, fairness, transparency, explainability, and trustworthiness.

Hatice Gunes: I am a computer scientist with over 20 years of experience at the intersection of machine learning, affective computing, and robotics, with applications in human-centred AI for adults, children, elderly, and people with disabilities. I have served as PI and co-I on major multidisciplinary research projects focused on AI and robotics for wellbeing. Over the past 5 years, I have conducted research on fairness and explainability, key areas within AI ethics.

A Appendix

This appendix contains Tables 1 and 2, which present quotes from participants from the three studies. The studies are differentiated as "Group 1" (members of the public at a science festival), "Group 2" (women in computer science), and "Group 3" (academics interested in the history and philosophy of science). The quotes were collected from the filled in Social Robot Co-Design Ethics canvas. The quotes have been categorised into the four themes, identified as ethical and socio-technical questions.

Topic	Quotes from participants
Theme 1: Is it safe? (Group 1)	G_1P_{05}: "Iterative designs and quality control, robots to spend eg 1 month off so can be tested for inappropriate learned behaviors" G_1P_{06}: "Getting attacked by the robot" G_1P_{06}: "The person putting too much trust in the robot, giving away too much personal information" G_1P_{05}: "Yes, because that affection won't be returned" G_1P_{01}: "Critical attachment period in young children? Should not be used as substitutes for parents away for long periods." G_1P_{04}: "Have robot's code be open source + verifiable" G_1P_{04}: "Just don't use CV algorithms. Test extensively on cross-section of people."
(Group 2)	G_2P_{12}: "Put safety requirement into the robot system design" G_2P_{01}: "The robot should be equipped with as many safety measures as possible to avoid and mitigate inappropriate behaviors. It is very important to design such safety measures" G_2P_{11}: "Robots being very close to user can harm the user" G_2P_{05}: "User get hurt. Children can damage the robot" G_2P_{01}: "The robot could incorrectly interact with its environment e.g fall off a table and break something" G_2P_{01}: "A user that gets upset may try to damage the robot" G_2P_{10}: "Isolation from community because of robot dependability"
(Group 3)	G_3P_{01}: "I guess they need to be tested with a wide range of people and very diverse." G_3P_{03}: "Unlikely problem in this robot scenario" G_3P_{01}: "If the robot is doing lots of movements it could hurt the user!" G_3P_{02}: "Could this relationship warp child development of attachment to humans?" G_3P_{01}: "In the context of wellbeing support, humans could feel like they would want the robot to recognize them or reciprocate the potential attachment. It might be weird to have to re-explain who you are to a robot that has supported you"
Theme 2: Who is the robot built for and with? (Group 1)	G_1P_{05} : "Users don't see themselves represented and/or the robot says insensitive things and/or can't respond appropriately to the user's needs" G_1P_{05} : "Robots should be designed by an array of people from different backgrounds and walks of life" G_1P_{04} : "CV algorithms could be biased by appearance" G_1P_{04} : "Just don't use CV algorithms. Test extensively on cross-section of people." G_1P_{01} : "Has human features, but remain androgynous and not race-specific"
(Group 2)	G_2P_{03}: "Feed robots with diverse data collected from different cultures" G_2P_{03}: "Provide a rich appropriate data to the robot DB/train data to teach robot what kind of acceptable behaviours" G_2P_{03}: "Treat users as white western users and ignore other types of users. Provide inappropriate advise/comments to some users from different culture. " G_2P_{01}: "If the robot should appeal to all groups, its dialogue should be relevant for all people otherwise, it can only talk about yoga and green juice. Price affects who it helps" G_2P_{02}: "Caregiver should play an important role in making them understand the role of the robot" G_2P_{01}: "Children or vulnerable groups who don't understand the interaction maybe should not use it"
(Group 3)	G_3P_{02}: "Is the robot's voice gendered? Racialized?" G_3P_{02}: "A user that is already experiencing "bullying" in everyday life can also experience it in the interaction with the robot." G_3P_{01}: "Sociocultural differences are very important in the kind of support needed. I wonder how robots can navigate that. " G_3P_{01}: "The robot's skills might need to be continuously improved and the robot could also be programmed to be very polite (?) Again, this is so culturally dependent (see: equality) that I'm not even sure if it's possible"

Table 1: **Theme 1 and 2:** Quotes from participants. $G_i, i = 1, 2, 3$ denotes the Group the participant took part in , and $P_j, j = 01, 02...$ denotes the ID number of the participant within that group.

Topic	Quotes from participants
Theme 3: Who owns the robot and the data? (Group 1)	G_1P_{05}: "Need laws and consideration from policy makers" G_1P_{01}: "Not an alternative to therapy, but if it does reach that point, shouldn't the same confidentiality concerns apply?" G_1P_{01}: "Only collecting non-sensitive data specific to the individual and maintain anonymity for more sensitive data? Consent forms?" G_1P_{02}: "Means of storage. Physical and accessibility" G_1P_{05}: "Need to consider confidentiality and GDPR and how it applies to a non-human subject that interacts like "it was human"" G_1P_{06}: "Making sure that the memory of the robot is deleted after the session"
(Group 2)	G_2P_{06}"Can the user really trust the robot if how their data is used is unclear" G_2P_{04}: "The data can be very personal, and if being used for improving the models, there's risk for privacy" G_2P_{06}: "Users on [unclear] like their privacy is violated or they feel the robot is alive [unclear] confidential even if they are aware data is stored" G_2P_{02}: "This can be a problem indeed because of for instance also in the case of positive psychology a user tells about [unclear] stories that could include sensitive data" G_2P_{01}: "User's wellbeing is sensitive, health information legally protected" G_2P_{07}: "Information collected by robot should only be accessible by people that are involved and users should be inform about it"
(Group 3)	G_3P_{04}: "Question of who owns the robots - employer? or potentially only your health care" G_3P_{02}: "How do users know that the robot has their best interests at heart?" G_3P_{01}: "Trust is clearly an issue in the context of people talking about their wellbeing. Not enough transparency could really hinder trust." G_3P_{01}: "The robot should be clean about the way it stacks data. Maybe it could disclose it at the beginning of session."
Theme 4: Why a robot? (Group 1)	G_1P_{04}: "Robot should be able to field questions about its intentions/abilities" G_1P_{04}: "Make robot's appearance/behavior deliberately unhuman-like" G_1P_{06}: "The person putting too much trust in the robot, giving away too much personal information" G_1P_{05}: "Yes, because that affection won't be returned"
(Group 2)	G_2P_{01}: "TV - transformers, Big Hero 6, Wall-E set unrealistic expectations of what the robot can feel/provide" G_2P_{01}: "Need clear instructions on who the robot is for, when prescribing them as treatment for people (for healthcare professionals). Setting realistic expectations for the user." G_2P_{06}: "There will be complexity in how to [unclear] that may be difficult to explain or clarify in terms most people understand" G_2P_{10}: "Robot should share its limitations when it's the case" G_2P_{11}: "Not every user likes robots! " G_2P_{09}: "Yes, [unclear] during a therapy session is necessary to get attached with your therapist otherwise no cure." G_2P_{09}: "Not using this with people with a really severe problem" G_2P_{05}: "Children can be attached to the robot as it was a friend" G_2P_{03}: I think yes, I don't think it is good for people to be attached to robots and forget how does it mean when you have a real human attachment"
(Group 3)	G_3P_{03}: "Need info about how robot is trained and how it is not trained" G_3P_{01}: "I think I would like the robot to share with me exactly who/what they are, how they work, what they can offer, what they cannot, what to expect, what the limits are, potential risks, etc, etc." G_3P_{01}: "User may need to be aware of how the robot may or may not recognize them and how it can or cannot interact with them its limits, etc, and the work it poses. " G_3P_{03}: "Need info about how robot is "The more the robot is "like" a person, the more these problems may emerge. But we want the robot to be "like a person" enough for wellness affects " G_3P_{02}: "This looks into the issue of ensuring users distinguish between humans + robots. Is it problematic to overemphasize robot empathy?"

Table 2: **Theme 3 and 4:** Quotes from participants. $G_i, i = 1, 2, 3$ denotes the Group the participant took part in , and $P_j, j = 01, 02...$ denotes the ID number of the participant within that group.

References

Arifah Addison, Christoph Bartneck, and Kumar Yogeeswaran. Robots can be more than black and white: examining racial bias towards robots. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 493–498, 2019.

Neziha Akalin, Pinar Uluer, and Hatice Kose. Non-verbal communication with a social robot peer: Towards robot assisted interactive sign language tutoring. In *2014 IEEE-RAS International Conference on Humanoid Robots*, pages 1122–1127. IEEE, 2014.

Malak Masnad Irshed Al-Qbilat. Accessibility requirements for human-robot interaction for socially assistive robots. *Ph. D., Universidad Carlos III de Madrid. Departamento de Informática*, 2022.

Sultan Almuaythir, Atul Kumar Singh, Mohammad Alhusban, and Ahmed Osama Daoud. Robotics technology: catalyst for sustainable development—impact on innovation, healthcare, inequality, and economic growth. *Discover Sustainability*, 5(1):486, 2024.

Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. Factsheets: Increasing trust in ai services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5):6–1, 2019.

Minja Axelsson. Social robot co-design canvases, Jul 2020. URL https://osf.io/jg2t8/. Open Science Framework, https://osf.io/jg2t8, Accessed: 11-08-2025.

Minja Axelsson, Mattia Racca, Daryl Weir, and Ville Kyrki. A participatory design process of a robotic tutor of assistive sign language for children with autism. In *2019 28th IEEE international conference on robot and human interactive communication (RO-MAN)*, pages 1–8. IEEE, 2019.

Minja Axelsson, Indu P Bodala, and Hatice Gunes. Participatory design of a robotic mental well-being coach. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pages 1081–1088. IEEE, 2021a.

Minja Axelsson, Raquel Oliveira, Mattia Racca, and Ville Kyrki. Social robot co-design canvases: A participatory design framework. *ACM Transactions on Human-Robot Interaction (THRI)*, 11(1):1–39, 2021b.

Minja Axelsson, Micol Spitale, and Hatice Gunes. Robots as mental well-being coaches: Design and ethical recommendations. *ACM Transactions on Human-Robot Interaction*, 2022.

Minja Axelsson, Micol Spitale, and Hatice Gunes. Robotic coaches delivering group mindfulness practice at a public cafe. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 86–90, 2023.

Minja Axelsson, Micol Spitale, and Hatice Gunes. " oh, sorry, i think i interrupted you": Designing repair strategies for robotic longitudinal well-being coaching. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 13–22, 2024.

Minja Axelsson, Nikhil Churamani, Atahan Çaldır, and Hatice Gunes. Participant perceptions of a robotic coach conducting positive psychology exercises: A qualitative analysis. *ACM Transactions on Human-Robot Interaction*, 14(2):1–27, 2025.

Kim Baraka, Patrícia Alves-Oliveira, and Tiago Ribeiro. An extended framework for characterizing social robots. *Human-robot interaction: evaluation methods and their standardization*, pages 21–64, 2020.

Julia Barnett and Nicholas Diakopoulos. Crowdsourcing impacts: exploring the utility of crowds for anticipating societal impacts of algorithmic decision making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 56–67, 2022.

Christoph Bartneck and Jun Hu. Exploring the abuse of robots. *Interaction Studies*, 9(3):415–433, 2008.

Stevie Bergman, Nahema Marchal, John Mellor, Shakir Mohamed, Iason Gabriel, and William Isaac. Stela: a community-centred approach to norm elicitation for ai alignment. *Scientific Reports*, 14(1):6616, 2024.

Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 648–657, 2020.

Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 401–413, 2021.

Miriam Bilac, Marine Chamoux, and Angelica Lim. Gaze and filled pause detection for smooth human-robot conversations. In *2017 IEEE-RAS 17th international conference on humanoid robotics (humanoids)*, pages 297–304. IEEE, 2017.

Abeba Birhane. Bending the arc of ai towards the public interest. `https://aial.ie/pages/aiparis/`, 2025. Accessed: 2025-03-25.

Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. Power to the people? opportunities and challenges for participatory ai. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–8, 2022.

Indu P Bodala, Nikhil Churamani, and Hatice Gunes. Creating a robot coach for mindfulness and wellbeing: A longitudinal study. *arXiv preprint arXiv:2006.05289*, 2020.

Andrea Bonaccorsi, Riccardo Apreda, and Gualtiero Fantoni. Expert biases in technology foresight. why they are a problem and how to mitigate them. *Technological Forecasting and Social Change*, 151:119855, 2020.

Ted K Bradshaw. The post-place community: Contributions to the debate about the definition of community. *Community development*, 39(1):5–16, 2008.

Philip Brey. Ethics of emerging technology. *The ethics of technology: Methods and approaches*, pages 175–191, 2017.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

Joseph Cameron, Jiaee Cheong, Micol Spitale, and Hatice Gunes. Multimodal gender fairness in depression prediction: Insights on data from the usa & china. *arXiv preprint arXiv:2408.04026*, 2024.

Xinyun Cao, Yunyi Wu, Mark Nielsen, and Fuxing Wang. Does appearance affect children's selective trust in robots' social and emotional testimony? *Journal of Applied Developmental Psychology*, 96:101739, 2025.

Tara Capel and Margot Brereton. What is human-centered about human-centered ai? a map of the research landscape. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–23, 2023.

Felix Carros, Tobias Störzinger, Anne Wierling, Adrian Preussner, and Peter Tolmie. Ethical, legal & participatory concerns in the development of human-robot interaction: lessons from eight research projects with social robots in real-world scenarios. *i-com*, 21(2):299–309, 2022.

Stephen Cave and Kanta Dihal. The whiteness of ai. *Philosophy & Technology*, 33(4):685–703, 2020.

Yaqub Chaudhary and Jonnie Penn. Beware the intention economy: Collection and commodification of intent via large language models. *Harvard Data Science Review*, (Special Issue 5), 2024.

Jiaee Cheong, Sinan Kalkan, and Hatice Gunes. The hitchhiker's guide to bias and fairness in facial affective signal processing: Overview and techniques. *IEEE Signal Processing Magazine*, 38(6):39–49, 2021.

Jiaee Cheong, Sinan Kalkan, and Hatice Gunes. Causal structure learning of bias for fair affect recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 340–349, 2023a.

Jiaee Cheong, Selim Kuzucu, Sinan Kalkan, and Hatice Gunes. Towards gender fairness for mental health prediction. In *IJCAI*, pages 5932–5940, 2023b.

Jiaee Cheong, Micol Spitale, and Hatice Gunes. "it's not fair!" – fairness for a small dataset of multi-modal dyadic mental well-being coaching. In *ACII 2023*, 2023c.

Jiaee Cheong, Sinan Kalkan, and Hatice Gunes. Fairrefuse: referee-guided fusion for multimodal causal fairness in depression detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 7224–7232, 2024a.

Jiaee Cheong, Micol Spitale, and Hatice Gunes. Small but fair! fairness for multimodal human-human and robot-human mental wellbeing coaching. *arXiv preprint arXiv:2407.01562*, 2024b.

Jiaee Cheong, Aditya Bangar, Sinan Kalkan, and Hatice Gunes. U-fair: Uncertainty-based multimodal multitask learning for fairer depression detection. In *Proceedings of the 4th Machine Learning for Health Symposium*, volume 259 of *Proceedings of Machine Learning Research*, pages 203–218. PMLR, 15–16 Dec 2025.

Charlene H Chu, Rune Nyrup, Kathleen Leslie, Jiamin Shi, Andria Bianchi, Alexandra Lyn, Molly McNicholl, Shehroz Khan, Samira Rahimi, and Amanda Grenier. Digital ageism: challenges and opportunities in artificial intelligence for older adults. *The Gerontologist*, 62(7):947–955, 2022.

Jon Chun and Katherine Elkins. The crisis of artificial intelligence: a new digital humanities curriculum for human-centred ai. *International Journal of Humanities and Arts Computing*, 17(2):147–167, 2023.

Nikhil Churamani, Jiaee Cheong, Sinan Kalkan, and Hatice Gunes. Towards causal replay for knowledge rehearsal in continual learning. In *AAAI Bridge Program on Continual Causality*, pages 63–70. PMLR, 2023.

Victoria Clarke and Virginia Braun. Thematic analysis. *The journal of positive psychology*, 12(3):297–298, 2017.

Houston Claure, Mai Lee Chang, Seyun Kim, Daniel Omeiza, Martim Brandao, Min Kyung Lee, and Malte Jung. Fairness and transparency in human-robot interaction. In *2022 HRI*, pages 1244–1246. IEEE, 2022.

EQUALS Skills Coalition, Mark West, Rebecca Kraut, Chew Han Ei, et al. I'd blush if i could: closing gender divides in digital skills through education, 2019.

Sasha Costanza-Chock. *Design justice: Community-led practices to build the worlds we need.* The MIT Press, 2020.

Kate Crawford. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press, 2021.

Clara Crivellaros, Lizzie Coles-Kemp, Alan Dix, and Ann Light. Co-creating conditions for social justice in digital societies: modes of resistance in hci collaborative endeavors and evolving socio-technical landscapes. *ACM Transactions on Computer-Human Interaction*, 2025.

Luisa Damiano and Paul Dumouchel. Anthropomorphism in human–robot co-evolution. *Frontiers in psychology*, 9: 468, 2018.

DartmouthNews. First therapy chatbot trial yields mental health benefits. `https://home.dartmouth.edu/news/2025/03/first-therapy-chatbot-trial-yields-mental-health-benefits`, 2025. Accessed: 2025-04-01.

Julian De Freitas and I Glenn Cohen. The health risks of generative ai-based wellness apps. *Nature medicine*, 30(5): 1269–1275, 2024.

DemocracyNext. A tech-enhanced student assembly. `https://studentassembly.mit.edu/`, 2024. Accessed: 2025-03-25.

Nathaniel S Dennler, Mina Kian, Stefanos Nikolaidis, and Maja Matarić. Designing robot identity: The role of voice, clothing, and task on robot gender perception. *arXiv preprint arXiv:2404.00494*, 2024.

Anthony Dunne and Fiona Raby. *Speculative Everything, With a new preface by the authors: Design, Fiction, and Social Dreaming*. MIT press, 2024.

Leonard E Egede. Race, ethnicity, culture, and disparities in health care. *Journal of general internal medicine*, 21(6): 667, 2006.

EU. Complete guide to gdpr compliance. `https://gdpr.eu/`, 2025. European Union, Accessed: 2025-05-15.

Wendy Faulkner. 'nuts and bolts and people' gender troubled engineering identities. *Engineering Identities, Epistemologies and Values: Engineering Education and Practice in Context, Volume 2*, pages 23–40, 2015.

Julia Fink. Anthropomorphism and human likeness in the design of robots and human-robot interaction. In *Social Robotics: 4th International Conference, ICSR 2012, Chengdu, China, October 29-31, 2012. Proceedings 4*, pages 199–208. Springer, 2012.

Patrick Fournier. *When citizens decide: Lessons from citizen assemblies on electoral reform*. Oxford University Press, 2011.

Iason Gabriel and Vafa Ghazavi. The challenge of value alignment. In *The Oxford handbook of digital ethics*. Oxford University Press Oxford, 2022.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Greg Guest, Arwen Bunce, and Laura Johnson. How many interviews are enough? an experiment with data saturation and variability. *Field methods*, 18(1):59–82, 2006.

Gordon C Nagayama Hall, Elliot T Berkman, Nolan W Zane, Frederick TL Leong, Wei-Chin Hwang, Arthur M Nezu, Christine Maguth Nezu, Janie J Hong, Joyce P Chu, and Ellen R Huang. Reducing mental health disparities by increasing the personal relevance of interventions. *American Psychologist*, 76(1):91, 2021.

Paula Hall and Debbie Ellis. A systematic review of socio-technical gender bias in ai algorithms. *Online Information Review*, 47(7):1264–1279, 2023.

Lelia Marie Hampton. Black feminist musings on algorithmic oppression. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 1, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi:10.1145/3442188.3445929. URL `https://doi.org/10.1145/3442188.3445929`.

Sven Ove Hansson. *The ethics of technology: methods and approaches*. Rowman & Littlefield, 2017.

Kerstin S Haring, Katsumi Watanabe, Mari Velonaki, Chad C Tossell, and Victor Finomore. Ffab—the form function attribution bias in human–robot interaction. *IEEE Transactions on Cognitive and Developmental Systems*, 10(4): 843–851, 2018.

Rakibul Hasan, Arto Ojala, Sara Quach, Park Thaichon, and Scott Weaven. The dark side of ai anthropomorphism: A case of misplaced trustworthiness in service provisions. In *Proceedings of the 58th Hawaii International Conference on System Sciences*. University of Hawaii, 2025.

Yugo Hayashi, Keita Kiuchi, Shigen Shimojo, Lisa Abe, and Emika Watanabe. A comparative study of older and younger adults using solution-focused brief therapy with an active listening counseling robot. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction*, pages 1337–1341, 2025.

Xiaoxuan Hei, Chuang Yu, Heng Zhang, and Adriana Tapus. A bilingual social robot with sign language and natural language. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 526–529, 2024.

Michael V Heinz, Daniel M Mackin, Brianna M Trudeau, Sukanya Bhattacharya, Yinzhou Wang, Haley A Banta, Abi D Jewett, Abigail J Salzhauer, Tess Z Griffin, and Nicholas C Jacobson. Randomized trial of a generative ai chatbot for mental health treatment. *NEJM AI*, 2(4):AIoa2400802, 2025.

Lukas Hindemith, Christiane B Wiebel-Herboth, Heiko Wersing, Britta Wrede, and Anna-Lisa Vollmer. Improving hri through robot architecture transparency. *International Journal of Social Robotics*, pages 1–21, 2025.

Tom Hitron, Benny Megidish, Etay Todress, Noa Morag, and Hadas Erel. Ai bias in human-robot interaction: An evaluation of the risk in gender biased robots. In *2022 RO-MAN*. IEEE, 2022.

Colin Holbrook, Umesh Krishnamurthy, Paul P Maglio, and Alan R Wagner. Physical anthropomorphism (but not gender presentation) influences trust in household robots. *Computers in Human Behavior: Artificial Humans*, 3: 100114, 2025.

Yoyo Tsung-Yu Hou, EunJeong Cheon, and Malte F Jung. Power in human-robot interaction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 269–282, 2024.

Bronwyn Howell. Weird? institutions and consumers' perceptions of artificial intelligence in 31 countries. *AI & SOCIETY*, pages 1–23, 2025.

Shin-Shin Hua and Haydn Belfield. Effective enforceability of eu competition law under ai development scenarios: a framework for anticipatory governance. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 596–605, 2023.

Sooyeon Jeong, Sharifa Alghowinem, Laura Aymerich-Franch, Kika Arias, Agata Lapedriza, Rosalind Picard, Hae Won Park, and Cynthia Breazeal. A robotic positive psychology coach to improve college students' wellbeing. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 187–194. IEEE, 2020.

Sooyeon Jeong, Laura Aymerich-Franch, Kika Arias, Sharifa Alghowinem, Agata Lapedriza, Rosalind Picard, Hae Won Park, and Cynthia Breazeal. Deploying a robotic positive psychology coach to improve college students' psychological well-being. *User Modeling and User-Adapted Interaction*, 33(2):571–615, 2023.

Michelle J Johnson, Megan A Johnson, Justine S Sefcik, Pamela Z Cacchione, Caio Mucchiani, Tessa Lau, and Mark Yim. Task and design requirements for an affordable mobile service robot for elder care in an all-inclusive care for elders assisted-living setting. *International Journal of Social Robotics*, 12:989–1008, 2020.

Katarzyna Kabacińska, Tony J Prescott, and Julie M Robillard. Socially assistive robots as mental health interventions for children: a scoping review. *International Journal of Social Robotics*, 13(5):919–935, 2021.

Pratyusha Kalluri. Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature*, 583(7815):169–169, 2020.

Tae Wan Kim, John Hooker, and Thomas Donaldson. Taking principles seriously: A hybrid approach to value alignment in artificial intelligence. *Journal of Artificial Intelligence Research*, 70:871–890, 2021.

Tae Woo Kim, Li Jiang, Adam Duhachek, Hyejin Lee, and Aaron Garvey. Do you mind if i ask you a personal question? how ai service agents alter consumer self-disclosure. *Journal of Service Research*, 25(4):649–666, 2022.

Taewook Kim, Hyeok Kim, Angela C Roberts, Maia Jacobs, and Matthew Kay. Opportunities in mental health support for informal dementia caregivers suffering from verbal agitation. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–26, 2024.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

Selim Kuzucu, Jiaee Cheong, Hatice Gunes, and Sinan Kalkan. Uncertainty as a fairness measure. *Journal of Artificial Intelligence Research*, 81:307–335, 2024.

Angus Man Ho Kwok, Jiaee Cheong, Sinan Kalkan, and Hatice Gunes. Machine learning fairness for depression detection using eeg data. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2025.

Guy Laban, Arvid Kappas, Val Morrison, and Emily S Cross. Opening up to social robots: how emotions drive self-disclosure behavior. In *2023 32nd ieee international conference on robot and human interactive communication (ro-man)*, pages 1697–1704. IEEE, 2023.

Clara Lachemaier, Eleonore Lumer, Hendrik Buschmeier, and Sina Zarrieß. Towards understanding the entanglement of human stereotypes and system biases in human-robot interaction. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 646–649, 2024.

Matthew Lamsey, Meredith D Wells, Lydia Hamby, Paige Scanlon, Rouida Siddiqui, You Liang Tan, Jerry Feldman, Charles C Kemp, and Madeleine E Hackney. Exercise specialists evaluation of robot-led physical therapy for people with parkinsons disease. *arXiv preprint arXiv:2502.04635*, 2025.

Tuukka Lehtiniemi and Jesse Haapoja. Data agency at stake: Mydata activism and alternative frames of equal participation. *new media & society*, 22(1):87–104, 2020.

Jie Li, Junpei Zhong, and Ning Wang. A multimodal human-robot sign language interaction framework applied in social robots. *Frontiers in neuroscience*, 17:1168888, 2023.

Yueqi Li and Sanjay Goel. Making it possible for the auditing of ai: A systematic review of ai audits and ai auditability. *Information Systems Frontiers*, pages 1–31, 2024.

Daria Loi, Christine T Wolf, Jeanette L Blomberg, Raphael Arar, and Margot Brereton. Co-designing ai futures: Integrating ai ethics, social computing, and design. In *Companion publication of the 2019 on designing interactive systems conference 2019 companion*, pages 381–384, 2019.

Hamza Mahdi, Sami Alperen Akgun, Shahed Saleh, and Kerstin Dautenhahn. A survey on the design and evolution of social robots—past, present and future. *Robotics and Autonomous Systems*, 156:104193, 2022.

Martin N Marshall. Sampling for qualitative research. *Family practice*, 13(6):522–526, 1996.

Kayla Matheus, Marynel Vázquez, and Brian Scassellati. Ommie: The design and development of a social robot for anxiety reduction. *ACM Transactions on Human-Robot Interaction*, 14(2):1–34, 2025.

Richard Milne, Catherine Galloway, Mariam Rashid, Daniela Boraschi, Claudette Burch, and Anna Middleton. The hopes and fears lab: enabling dialogue on discovery science. *Journal of Science Communication*, 2024.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

Jakob Mökander. Auditing of ai: Legal, ethical and technical approaches. *Digital Society*, 2(3):49, 2023.

António B Moniz and Bettina-Johanna Krings. Robots working with humans or humans working with robots? searching for social dimensions in new human-robot interaction in industry. *Societies*, 6(3):23, 2016.

Tatsuya Nomura, Takayuki Kanda, Tomohiro Suzuki, and Sachie Yamada. Do people with social anxiety feel anxious about interacting with a robot? *Ai & Society*, 35:381–390, 2020.

Don Norman. What are complex socio-technical systems? `https://www.interaction-design.org/literature/topics/complex-socio-technical-systems`, 2021. Accessed: 2025-03-25.

Rune Nyrup. From general principles to procedural values: responsible digital health meets public health ethics. *Frontiers in Digital Health*, 3:690417, 2021.

Rune Nyrup, Charlene H. Chu, and Elena Falco. 309digital ageism, algorithmic bias, and feminist critical theory. In *Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines*. Oxford University Press, 10 2023. ISBN 9780192889898. doi:10.1093/oso/9780192889898.003.0018. URL `https://doi.org/10.1093/oso/9780192889898.003.0018`.

Tobi Ogunyale, De'Aira Bryant, and Ayanna Howard. Does removing stereotype priming remove bias? a pilot human-robot interaction study. *arXiv preprint arXiv:1807.00948*, 2018.

Raquel Oliveira and Elmira Yadollahi. Robots in movies: A content analysis of the portrayal of fictional social robots. *Behaviour & Information Technology*, 43(5):970–987, 2024.

Anastasia K Ostrowski. *How Do We Design Robots Equitably? Engaging Design Justice, Design Fictions, and Co-Design in Human-Robot Interaction Design and Policymaking Processes*. PhD thesis, Massachusetts Institute of Technology, 2023.

Anastasia K Ostrowski, Raechel Walker, Madhurima Das, Maria Yang, Cynthia Breazea, Hae Won Park, and Aditi Verma. Ethics, equity, & justice in human-robot interaction: A review and future directions. In *2022 RO-MAN*, pages 969–976. IEEE, 2022.

Almudena Otegui Carles, José Antonio Fraiz Brea, and Noelia Araújo Vila. Gender stereotypes in robotics in the field of tourism and hospitality. a conceptual paper. *Journal of Hospitality and Tourism Technology*, 16(2):389–408, 2025.

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

Giulia Perugia, Stefano Guidi, Margherita Bicchi, and Oronzo Parlangeli. The shape of our bias: Perceived age and gender in the humanoid robots of the abot database. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 110–119. IEEE, 2022.

Antti Poikola, Kai Kuikkaniemi, Ossi Kuittinen, Harri Honko, Aleksi Knuutila, and Viivi Lähteenoja. Mydata–an introduction to human-centric use of personal data (3rd. edition). `https://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/162405/MyData%20-%20introduction%20to%20human-centric%20use%20of%20personal%20data%203rd%20revised%20edition.pdf?sequence=1`, 2020. Accessed: 2025-08-11.

Hamed Pourfannan, Rachel Young, and Alessandro Di Nuovo. Integrating humanoid robots in stroke rehabilitation: Practitioners' expectations and insights. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction*, pages 1553–1557, 2025.

Christina Pozo, Charles S Carver, A Rodney Weflens, and Michael F Scheier. Social anxiety and social perception: Construing others' reactions to the self. *Personality and Social Psychology Bulletin*, 17(4):355–362, 1991.

Vinodkumar Prabhakaran, Rida Qadri, and Ben Hutchinson. Cultural incongruencies in artificial intelligence. *arXiv preprint arXiv:2211.13069*, 2022.

Malak Qbilat, Ana Iglesias, and Tony Belpaeme. A proposal of accessibility guidelines for human-robot interaction. *Electronics*, 10(5):561, 2021.

Inioluwa Deborah Raji and Jingying Yang. About ml: Annotation and benchmarking on understanding and transparency of machine learning lifecycles. *arXiv preprint arXiv:1912.06166*, 2019.

Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. Where responsible ai meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–23, 2021.

Bogdana Rakova, Renee Shelby, and Megan Ma. Terms-we-serve-with: Five dimensions for anticipating and repairing algorithmic harm. *Big Data & Society*, 10(2):20539517231211553, 2023.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Rosa Romano. Ethical issues on artificial intelligence and human relationships. In *INTED2024 Proceedings*, pages 6902–6909. IATED, 2024.

Selma Šabanović, Vicky Charisi, Tony Belpaeme, Cindy L Bethel, Maja Matarić, Robin Murphy, and Shelly Levy-Tzedek. "robots for good": Ten defining questions. *Science Robotics*, 8(84):eadl4238, 2023.

Antonio Saporito, Parinaz Tabari, Mattia De Rosa, Vittorio Fuccella, and Gennaro Costagliola. Exploring the privacy horizons: A survey on hci & hri. In *International Conference on Computational Science and Its Applications*, pages 113–125. Springer, 2024.

Batuhan Sayis and Hatice Gunes. Technology-assisted journal writing for improving student mental wellbeing: Humanoid robot vs. voice assistant. In *Companion of the 2024 acm/ieee international conference on human-robot interaction*, pages 945–949, 2024.

Peter Schaar. Privacy by design. *Identity in the Information Society*, 3(2):267–274, 2010.

Arielle AJ Scoglio, Erin D Reilly, Jay A Gorman, and Charles E Drebing. Use of social robots in mental health and well-being research: systematic review. *Journal of medical Internet research*, 21(7):e13322, 2019.

Sofia Serholt, Sara Ljungblad, and Niamh Ní Bhroin. Introduction: special issue—critical robotics research. *AI & SOCIETY*, 37(2):417–423, 2022.

Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 723–741, 2023.

Joshua Skewes, David M Amodio, and Johanna Seibt. Social robotics and the modulation of social perception and bias. *Philosophical Transactions of the Royal Society B*, 374(1771):20180037, 2019.

David Smith and Marilyn Fitzpatrick. Patient-therapist boundary issues: An integrative review of theory and research. *Professional psychology: research and practice*, 26(5):499, 1995.

Ewan Soubutts, Pranita Shrestha, Brittany I Davidson, Chengcheng Qu, Charlotte Mindel, Aaron Sefi, Paul Marshall, and Roisin McNaney. Challenges and opportunities for the design of inclusive digital mental health tools: Understanding culturally diverse young people's experiences. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2024.

Robert Sparrow. Do robots have race?: Race, social construction, and hri. *IEEE Robotics & Automation Magazine*, 27 (3):144–150, 2019.

Micol Spitale and Hatice Gunes. Affective robotics for wellbeing: A scoping review. In *2022 10th International conference on affective computing and intelligent interaction workshops and demos (ACIIW)*, pages 1–8. IEEE, 2022.

Micol Spitale, Minja Axelsson, and Hatice Gunes. Robotic mental well-being coaches for the workplace: An in-the-wild study on form. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 301–310, 2023a.

Micol Spitale, Minja Axelsson, Neval Kara, and Hatice Gunes. Longitudinal evolution of coachees' behavioural responses to interaction ruptures in robotic positive psychology coaching. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 315–322. IEEE, 2023b.

Micol Spitale, Minja Axelsson, and Hatice Gunes. Appropriateness of llm-equipped robotic well-being coach language in the workplace: A qualitative evaluation. *arXiv preprint arXiv:2401.14935*, 2024a.

Micol Spitale, Minja Axelsson, Sooyeon Jeong, Paige Tuttoşı, Caitlin A Stamatis, Guy Laban, Angelica Lim, and Hatice Gunes. Past, present, and future: A survey of the evolution of affective robotics for well-being. *arXiv preprint arXiv:2407.02957*, 2024b.

Micol Spitale, Jiaee Cheong, and Hatice Gunes. Underneath the numbers: Quantitative and qualitative gender fairness in llms for depression prediction. *arXiv preprint arXiv:2406.08183*, 2024c.

Megan Strait, Ana Sánchez Ramos, Virginia Contreras, and Noemi Garcia. Robots racialized in the likeness of marginalized social identities are subject to greater dehumanization than those racialized as white. In *2018 27th IEEE international symposium on robot and human interactive communication (RO-MAN)*, pages 452–457. IEEE, 2018.

Michael Suguitan and Guy Hoffman. Blossom: A handcrafted open-source robot. *ACM Transactions on Human-Robot Interaction (THRI)*, 8(1):1–27, 2019.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

Sita M Syal and Julia Kramer. Design and justice: A scoping review in engineering design. *Journal of Mechanical Design*, 147(5), 2025.

Sherry Turkle. A nascent robotics culture: New complicities for companionship. In *Machine ethics and robot ethics*, pages 107–116. Routledge, 2020.

UKCP. Ukcp code of ethics and professional practice. `https://www.psychotherapy.org.uk/media/bkjdm33f/ukcp-code-of-ethics-and-professional-practice-2019.pdf`, 2019. Accessed: 2025-04-01.

UKGovernment. Regulating medical devices in the uk. `https://www.gov.uk/guidance/regulating-medical-devices-in-the-uk`, 2025. Accessed: 2025-04-01.

Niels van Berkel, Benjamin Tag, Jorge Goncalves, and Simo Hosio. Human-centred artificial intelligence: a contextual morality perspective. *Behaviour & Information Technology*, 41(3):502–518, 2022.

Carissa Véliz. *Privacy is power*. Melville House Brooklyn, 2021.

Tosan Velor. *A Low-cost Social Companion Robot for Children with Autism Spectrum Disorder*. PhD thesis, Université d'Ottawa/University of Ottawa, 2020.

Alfio Ventura, Christopher Starke, Francesca Righetti, and Nils Köbis. Relationships in the age of ai: A review on the opportunities and risks of synthetic relationships to reduce loneliness, Mar 2025. URL `osf.io/preprints/psyarxiv/w7nmz_v1`.

Eduard Fosch Villaronga, Peter Kieseberg, and Tiffany Li. Humans forget, machines remember: Artificial intelligence and the right to be forgotten. *Computer Law & Security Review*, 34(2):304–313, 2018.

Adrian Weller. Transparency: motivations and challenges. In *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 23–40. Springer, 2019.

B Whitworth and A Ahmad. Socio-technical system design, 2014. URL `https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/socio-technical-system-design`.

Alan FT Winfield and Marina Jirotka. The case for an ethical black box. In *Towards Autonomous Robotic Systems: 18th Annual Conference, TAROS 2017, Guildford, UK, July 19–21, 2017, Proceedings 18*, pages 262–273. Springer, 2017.

Alan FT Winfield and Matthew Studley. On the relationship between benchmarking, standards and certification in robotics and ai. *arXiv preprint arXiv:2309.12139*, 2023.

Katie Winkle, Ryan Blake Jackson, Gaspar Isaac Melsión, Dražen Brščić, Iolanda Leite, and Tom Williams. Norm-breaking responses to sexist abuse: A cross-cultural human robot interaction study. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 120–129. IEEE, 2022.

Katie Winkle, Donald McMillan, Maria Arnelid, Katherine Harrison, Madeline Balaam, Ericka Johnson, and Iolanda Leite. Feminist human-robot interaction: Disentangling power, principles and practice for better, more ethical hri. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 72–82, 2023.

Baijun Xie and Chung Hyuk Park. An empathetic social robot with modular anxiety interventions for autistic adolescents. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, pages 1148–1155. IEEE, 2024.

Yifei Zhu, Ruchen Wen, and Tom Williams. Robots for social justice (r4sj): Toward a more equitable practice of human-robot interaction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 850–859, 2024.

Shoshana Zuboff. The age of surveillance capitalism. In *Social theory re-wired*, pages 203–213. Routledge, 2023.

Theresa Züger and Hadi Asghari. Ai for the public. how public interest theory shifts the discourse on ai. *AI & SOCIETY*, 38(2):815–828, 2023.