Assessing and Addressing Ethical Risk from Anthropomorphism and Deception in Socially Assistive Robots

Katie Winkle
KTH Royal Institute of Technology
Stockholm, Sweden
winkle@kth.se

Praminda Caleb-Solly
Bristol Robotics Laboratory
Bristol, U.K.
Praminda.Caleb-solly@uwe.ac.uk

Ute Leonards
School of Psychological Science
University of Bristol, Bristol, U.K.
ute.leonards@bristol.ac.uk

Ailie Turton
Bristol Robotics Laboratory
Bristol, U.K.
Ailie.Turton@uwe.ac.uk

Paul Bremner
Bristol Robotics Laboratory
Bristol, U.K.
paul.bremner@brl.ac.uk

ABSTRACT

In this paper we apply the recent concept of robot Ethical Risk Assessment to an exemplar Socially Assistive Robot (SAR); specifically considering ethical risks posed by anthropomorphism in this context. We draw on two complimentary studies to demonstrate that anthropomorphism is important to overall SAR function and overall relatively low ethical risk. As such, rather than avoiding anthropomoprhism all together (as suggested in a recently published standard on robot ethics), we suggest anthropomorphism in SARs should be a customisable trait that can be adapted to the user.

KEYWORDS

socially assistive robots, anthropomorphism, ethics, responsible robotics

ACM Reference Format:
Katie Winkle, Praminda Caleb-Solly, Ute Leonards, Ailie Turton, and Paul Bremner. 2021. Assessing and Addressing Ethical Risk from Anthropomorphism and Deception in Socially Assistive Robots. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21), March 8–11, 2021, Boulder, CO, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3434073.3444666

1 INTRODUCTION

By definition, Socially Assistive Robots (SARs) provide assistance through their social interaction [4]; typically being used to change user behaviour through social influence effects. Example applications include health/fitness [7, 10, 16] and education [9, 14], for which the SAR is normally designed to emulate the role of a human. This could involve the SAR acting as a peer (e.g. a learning companion [9]) or instead acting as an authority figure (e.g. a fitness instructor [16]). It is perhaps unsurprising then that SARs are typically quite anthropomorphic, as they are designed to leverage the same interaction cues and motivational strategies seen in Human-Human Interaction (HHI).

However, there has always been some argument as to whether anthropomorphism is really a desirable trait in (assistive) HRI (see e.g. [17] for a review of ethics literature regarding the use of robots in care). Specifically, it has been suggested that anthropomorphic behaviours are deceptive, because they essentially suggest a level of agency and social, affective capabilities that aren't *actually* present [3, 15]. This issue has received more attention lately as there has been an increasing focus on the topic of ethics in AI and robotics more broadly. A recent review identified 24 distinct sets of ethical principles for robotics and AI[1] to have emerged since 2009. Another output of this effort is the world's first explicitly ethical standard in robotics: BS8611-2016 *Guide to the ethical design and application of robots and robotic systems* [2].

1.1 Ethical Risk Assessment

Winfield and Winkle recently used BS8611 to define the concept of Ethical Risk Assessment (ERA) in the context of responsible robotics [18]. ERA represents a *practical* tool for systematically assessing and mitigating the ethical risk that a particular robot might pose. In this context, BS8611 aims to provide guidance on what ethical risks designers might look for when undertaking an ERA.

BS8611 is designed for application to a range of domains, and many of the ethical hazards identified are generally applicable to any robot. For example, issues concerning wastage and destruction of the environment arise from the manufacture, use and eventual decommissioning of any robot system. However, reviewing BS8611 quickly identifies a number of ethical risks that are *particularly* pertinent to social (assistive) robots. Specifically, the standard identifies the risks of *anthropomorphization* and *deception*. It is suggested that designers should avoid *"unnecessary anthropomorphization"* and *"deception due to the behaviour and/or appearance of the robot and ensure transparency of its robotic nature"* (page 3 of [2]). In both cases, *user validation* and *expert guidance* are given as tools for verification/validation for assessing and mitigating such risks.

As noted above, the suggestion that deception and anthropomorphization pose ethical risks for HRI is not novel; but only very recent work has specifically considered practical ERA of a social robot.

[1] http://alanwinfield.blogspot.com/2019/04/an-updated-round-up-of-ethical.html

HRI '21, March 8–11, 2021, Boulder, CO, USA

ACM ISBN 978-1-4503-8289-2/21/03…$15.00
https://doi.org/10.1145/3434073.3444666

[18] documented a case study ERA for a hypothetical social robot teddy bear. The authors specifically highlight *addiction*, *deception*, *over-trusting* and the *uncanny valley* as potential ethical risks most associated with the robot's anthropomorphic design/behaviour. However, the intentional use of anthropomorphic and (arguably) deceptive behaviours continues to be prolific and seemingly uncontroversial in social HRI (e.g. suggestion of emotional state [13] or robot 'gender' [8]). Further, with the exception of papers specifically considering ethical and philosophical arguments surrounding HRI behaviours (e.g. [17]) few works in social and/or socially assistive HRI make any reference to what ethical risk might result from the behaviours they describe. To the authors' knowledge, no previous works have specifically considered if/how (re-)designing such behaviours to minimise/mitigate ethical risk might impact on their potential effectiveness.

This work aims to address these issues directly, specifically by (i) attempting to identify and assess risk associated with those ethical hazards of *anthropomorphization* and *deception*, as identified by BS8611, for an exemplar SAR, (ii) investigate if/how (re-)designing SAR behaviour to reduce this risk might impact on SAR efficacy and (iii) reflect on the mitigation strategies suggested by BS8611 and the practical implications for SAR design.

Notably, BS8611 does not provide a definition of the deception it identifies as being an ethical hazard, nor any justification as to exactly why it might be ethically undesirable. So for this work, given:

(i) BS8611 associates deception with anthropomorphization, the 'simulation of human behaviour' and transparent regarding the robot's 'nature'
(ii) typical use of anthropomorphic robot behaviour design in the context of SARs
(iii) previous works considering whether robot emotion portrayal is deceptive [3] and suggesting that encouraging user-robot relationship development is immoral [15]

we interpret and define deceptive robot design/behaviour to be the *intentional* use of social behavioural cues that imply human-level social/affective capabilities, specifically those suggesting the robot is *emotionally invested* in its interaction with the user and the user's resultant behaviour.

To ground the work in a realistic use case, we specifically consider a robot fitness instructor, designed to guide and encourage users through prescribed exercise sessions, in line with our previous works.

2 ETHICAL RISK ASSESSMENT OF A ROBOT EXERCISE COACH

In a previous article, we demonstrated that persuasive dialogue strategies taken from HHI could increase the efficacy of a SAR for motivating exercise [19]. Specifically, having the robot:

(i) demonstrate some affective interest in the participant
(ii) indicate it shared the participant's opinions on exercising

resulted in participants undertaking more exercise than when the robot engaged in a 'socially neutral' control dialogue of the same length. In portraying these behaviours, the robot presented itself anthropomorphically as an independent social agent *interested in* and able to *empathise with* the user.

Table 1 identifies ethical risks that might result from these types of behaviour, and proposes mitigation strategies for reducing these risks. These specific risks and the associated mitigation strategies were derived from (i) the guidance provided in BS8611 and (ii) the previously referred to exemplar case study on ERA of a social robot [18]. Importantly, this does not represent a full ERA of the system. These risks are only those *particularly* related to the utilisation of anthropomorphic, socially persuasive behaviours, and were chosen in part to highlight the potential divide between typical practice in SAR and the recommendations made by BS8611. Whilst we consider the specific application of a robot coach for rehabilitative exercises, we suggest these risks would also hold for most other applications of SAR in which the robot is motivating user engagement with some undesirable task.

3 RESEARCH QUESTIONS

Having identified anthropomorphization and deception as ethical hazards that are (i) particularly pertinent to SARs and (ii) potentially necessary for SARs to function effectively, we address the following research questions:

RQ1 What evidence is there that typical (objectively effective) SAR behaviours pose those ethical risks listed in Table 1?
RQ2 How could those SAR behaviours be (re-)designed for reduced ethical risk (c.f. BS8611) and what impact might such re-design have on their efficacy?
RQ3 What are the resultant, practical implications for applying ethical risk assessment and mitigation to SAR design?

4 METHODOLOGY

To address the above research questions, we refer to two complimentary HRI studies as summarised in Table 2. In this work, we refer to Study 1 in the context of considering the ethical risk and acceptability of typical SAR behaviours. Specifically, we present additional results concerning perceptions of deception in and acceptability of those behaviours that we have previously demonstrated to be *objectively* effective [19], but that appear to go against the recommendations in BS8611. Study 2 is a novel study designed to (i) provide an initial demonstration of how the aforementioned behaviours might be re-designed for reduced ethical risk (c.f. [2]) and (ii) investigate what impact this might have on efficacy of the robot. Both studies were approved by the Faculty of Science ethics committee of the University.

4.1 Study 1: Ethical Risk for a Typical SAR

As described in [19], Study 1 was a between-subjects study in which participants were invited to undertake a mock exercise session with the Pepper robot[2]. The exercise task was open-ended, such that participants could stop exercising at any time. The two conditions most pertinent to this work are the Goodwill and Similarity conditions, both of which utilised anthropomorphic behaviours and yielded significantly more (voluntary) participant exercise repetitions when compared to the more 'socially neutral' control condition. In the

[2]https://www.softbankrobotics.com/emea/en/pepper

Table 1: Highlighted ethical risks (most associated with social and/or anthropomorphic behaviours) for a robot therapy coach designed to guide and encourage users through prescribed exercise sessions.

Hazard	Risk	Mitigation
Deception	User believes robot has feelings (that are affected by social interaction/exercise completion or lack thereof with the user).	Minimise use of affective social interaction that suggests robot 'emotional state' or social agency. Be upfront about robot's 'nature' [2].
Over-trusting	User and/or other responsible human(s) (e.g. therapist, carer) believe the robot to be more capable than it actually is at assessing and encouraging exercise.	Have robot clearly refer to appropriate human(s). Do not suggest unrealistic feedback capabilities. Make robot's capabilities (and limitations) clear.
Uncanny Valley (+/ User Dislike)	User is uncomfortable with (or, as an extension of the typical uncanny valley phenomena, simply *dislikes*) the social or anthropomorphic robot or its behaviours.	Minimise unnecessary social behaviour and/or anthropomorphic design cues.

Table 2: Overview of the two studies referred to in this article. The detailed experimental design for Study 1 and the results marked * are presented in [19]. Here, we refer back to those results specifically in the context of assessing ethical risk, and present additional results from that study concerning deception and acceptability. Study 2 represents a novel study designed to investigate the potential impact of (re-)designing the Study 1 behaviours for reduced ethical risk c.f. [2].

Study	Medium	Design	Manipulations	Selected Measures
1	Laboratory: Interactive Task + Post-hoc Interview	Between Subject	Persuasive (anthropomorphic) dialogue strategy	Number of exercise repetitions* Responsibility ascription* Deception and Acceptability
2	Online: Observing Pre-Recorded Videos of Robot-User Interactions	Within Subject	Anthropomorpshism (and hence ethical risk) in robot's dialogue	Credibility [6]; Likeability [1] Responsibility ascription Deception and Acceptability Preferred robot(s)

Goodwill condition, the robot demonstrated affective interest in the user by indicating e.g. it was pleased to meet them, excited to work with them and providing emotionally matched responses to their feelings about the exercise session. In the Similarity condition, the robot indicated it shared all of the user's exercising preferences, e.g. whether it is better to work out with others or alone.

Directly after exercising with the robot, participants completed a questionnaire containing a number of credibility and likeability measures as well as questions on deception and acceptability (detailed below under Section 4.3). They were then invited to take part in a brief, post-hoc, semi-structured interview, focused on exploring participants' answers to these questions on deception and acceptability. All 92 participants elected to take part in the interview, and the resultant transcripts were analysed using the Framework method, following published guidelines on the analysis of qualitative research [5]. Detailed participant demographic data are given in the supplementary appendix.

4.2 Study 2: Re-designing SAR Behaviours for Reduced Ethical Risk

Based on the results from Study 1, a three condition, within-subject, video based study was designed to demonstrate (i) how the *Goodwill* and *Similarity* behaviours from Study 1 might be re-designed according to the mitigation strategies presented in Table 1 and (ii) what impact this might have on their efficacy for motivating exercise engagement.

Participants were asked to watch three videos, each demonstrating a 'different version' of Pepper interacting with a 'patient' (actor). The context of the interaction and robot functionality were designed to mirror that of Study 1, i.e. with the robot being used to guide and motivate a user through some prescribed exercises. Figure 1 shows a snapshot from one of the videos; with the same scene set-up being used in all video clips. Each 'version' of the robot presented exercises designed to target arthritic pain in a different part of the body, and condition ordering was counterbalanced across participants.

Significant care was taken to ensure actor behaviour was consistent across videos, in order to limit what participants might deduce from the actor's behaviour. As can be seen in Figure 1, the video angle showed only the back of the actor's head (hence, no facial expressions) and the actor's audio responses to the robot were pre-recorded once to be used across all videos. All exercises, their descriptors and related information was taken from the same public health service[3] and Arthritis Research[4] self-help material consulted for Study 1. The videos were preceded by the following introduction:

"[Actor] suffers from arthritis and has been seeing a physiotherapist for help in alleviating her symptoms. Typically this involves the physiotherapist prescribing some daily exercises that [Actor] can do at home. Like many patients, [Actor] struggles with finding the motivation to do her exercises. In this study you will be shown three

[3]https://www.nhs.uk/conditions/tennis-elbow/
[4]https://www.versusarthritis.org/about-arthritis/conditions/elbow-pain/

different versions of a robot which could guide [Actor] through these daily exercises when her therapist can't be there. After each video you will be asked some questions about each individual version of the robot, and at the end you will be asked some questions comparing all three."

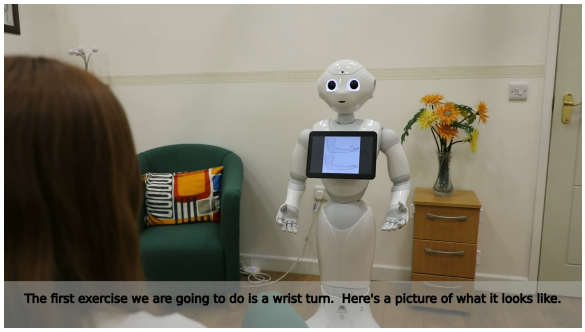

Figure 1: The socially assistive robot setting and scene setup used for all video clips in Study 2.

A total of 121 participants were recruited to the study representing 38 males, 82 females and 1 of undisclosed gender. Additional participant demographic data are given in the supplementary appendix. Participants were recruited through the Prolific online platform[5], through which they were reimbursed £2.50 (equivalent to national minimum wage) for their participation.

4.2.1 Experimental Conditions. The experimental conditions represent three different 'versions' of the Pepper robot, all of which fundamentally do the same thing: guide the user through a set of exercises. Two of the conditions are designed to demonstrate robot-participant goodwill and similarity through the robot's dialogue (as per Study 1), and the third condition is a control which utilises no such additional social interaction at all.

The two social conditions were then designed to vary the level of anthropomorphism (and hence, ethical risk) explicitly present in the robot's dialogue. The *Higher Risk* condition essentially replicates those Study 1 goodwill and similarity behaviours directly, i.e. with the robot presenting itself as an independent, highly social and life-like agent capable of affect and empathy. We suggest that this condition is fairly representative of SARs demonstrated in HRI literature, and meets the definition of deception we presented in Section 1 because the robot:

(i) suggests it is interested in/happy to get to know the user
(ii) attempts to demonstrate direct empathy with the user
(iii) suggests it is pleased by the user's performance.

The *Lower Risk* condition then represents our attempt to reduce ethical risk by re-designing the *Higher Risk* dialogue according the mitigation strategies presented in Table 1. This represents an attempt to have the robot still demonstrate some goodwill and similarity to the participant, but whilst also being *upfront about is robotic nature* as per the recommendations in BS8611 [2]. Specifically, compared to the *Higher Risk* condition, the *Lower Risk* robot:

(i) doesn't explicitly suggest it experiences human-like social/emotional feelings (e.g. being *pleased* to meet the user or *looking forward* to working together)
(ii) refers to being *built*, *programmed* and *designed*
(iii) doesn't attempt to empathise with the user *directly*, instead referring to similar difficulties faced by other humans
(iv) doesn't provide affective judgement directly (instead e.g. suggesting the patient's *therapist* would be impressed)

whilst still attempting to demonstrate some goodwill and similarity to the participant. Arguably this dialogue still implies some human-like social capability in the form of perspective taking; however the lack of suggested affective feeling by the robot itself still makes this less deceptive than our *Higher Risk* condition according to the definition of deception given in Section 1. The dialogue for each condition is given in the supplementary appendix and exemplar videos used for each condition can be found online [6]. All participants saw all three videos, with counterbalancing being used to avoid ordering effects.

4.3 Experimental Measures

The same measures were used across both studies (being adjusted as necessary for the switch from the in-person, between-subject Study 1 to the video-based, online, within-subject Study 2). A brief description of all measures is given below.

4.3.1 Credibility. Robot credibility was measured using questionnaire items designed to measure credibility of a human source; with 5-point Likert question items arranged in subscales of expertise, trustworthiness, goodwill and sociability (as presented in [6], adapted from [11] and [12]). Full question item descriptors are given in [19].

4.3.2 Likeability. Robot likeability was measured using the likeability scale of the Godspeed questionnaire [1] on a 5-point Likert scale. Other items from the Godspeed questionnaire were not included due to significant overlap with the credibility measure.

4.3.3 Ascription of Responsibility. Participants were asked to propose, if this robot were to be deployed for real world use in conjunction with a human therapist:

(i) how much responsibility *the robot* should be given for monitoring and advising the user
(ii) how much responsibility *the therapist* should be given for monitoring and advising the user

giving their responses on a 5-point Likert scale in line with the other measures.

4.3.4 Preferences. For (the within-subject) Study 2 only, participants were asked to identify:

(i) which of the robots they found most motivating
(ii) which of the robots they would prefer to work with

4.3.5 Deception and Acceptability. In Study 1, participants were asked in the post-exercise questionnaire whether they perceived the robot they had seen today to be deceptive, and whether that was acceptable. Participants were first given a brief explanation of *why*

[5]https://www.prolific.co/

[6]Higher Risk: https://youtu.be/G4k7Uxo4BYg; Lower Risk: https://youtu.be/b-nfUszHYqE; Control: https://youtu.be/WkSYI3cvohE

such behaviours might be considered deceptive, to account for their potential lack of experience/understanding regarding social robots and their capabilities. They were given the following answer options to choose from: *yes - deceptive and unacceptable, yes - deceptive but acceptable, not deceptive, not sure.* This question and their chosen answer was then re-visited during the post-hoc interview.

In Study 2, participants were asked whether any of the robots shown in the videos were deceptive, and were presented with the following answer options to choose from: *Version A (from video 1), Version B (from video 2), Version C (from video 3), None were deceptive* with the option of leaving a comment to explain their answer.

5 RESULTS

5.1 Deception and Acceptability

(RQ1) In Study 1, across all conditions, the majority of participants found the robot either *deceptive but acceptable* or *not deceptive*. Figure 2 shows that the spread of answers somewhat varies across the conditions as might be expected based on the behaviour manipulations they each represent. For example, the control robot was the only one to be rated as *not deceptive* by the majority (14/24) of participants. However, a significant number of participants (9/24) still found the robot to be deceptive (but acceptable). In addition, participants appeared to be more certain about the potential for deception in the *Similarity* condition when compared to the *Goodwill*, with an increased proportion of the Goodwill participants being *not sure* or finding the robot *not deceptive*.

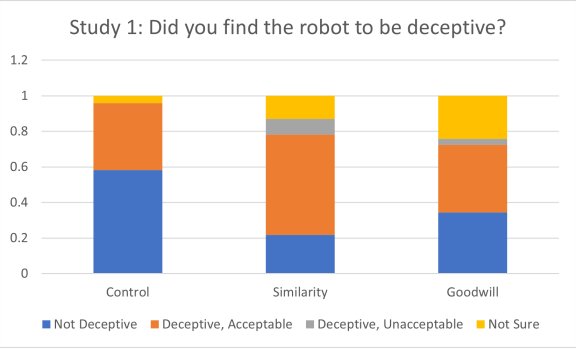

Figure 2: Frequency count of responses to the deception question in Study 1; normalised against the number of participants in each condition.

Open-ended questionnaire and interview data concerning participant reasoning on this question were coded for emergent themes regarding deception and acceptability. Common reasons given for the robot *not* being deceptive centered on its robotic nature; specifically that it was 'obviously' a robot and hence not deceiving anyone, or that as a robot it was incapable of deception and, related to that, that the robot was just following its programming:

[Goodwill7]: *"It's not deception because it's been programmed to do a certain thing, and so it's not deceiving anyone"*

[Similarity21] *"I did say it was deceptive on the form but because I feel like it's a program, a pre-programmed response. At the end of the day I realise it's a computer, well it's a robot and it's pre-programmed so to some degree it's deceptive... but I expect that"*

An unexpected theme however was the potential for deception if the robot suggested it was *watching or monitoring* the participant's exercise behaviour when in reality it wasn't.

[S13]: *"I put deceptive and unacceptable...you start off really friendly but then when I do the exercise wrong you don't tell me it's wrong so that's why I think it felt deceptive because it didn't seem to have my best interests at heart"*

[G27]: *"A real doctor would... maybe if you winced or something and then that real person would flag it, but if the robot does not have that kind of sensing capability ...so I doubt it is intentionally deceptive but the feedback it provides can be."*

Across all conditions and almost all participants, it was identified that the behaviours demonstrated were appropriate for the proposed application, and for making the robot more effective or usable in that context, therefore making them acceptable even if deceptive.

[S10]: *"If they are saying the same answers as you to encourage you I don't see that as being...I think they're being more helpful or you know someone you can relate to as you would do in human-human interaction you sometimes might feel more comfortable doing things around people with similar ideas to you"*

[G1]: *"If the reason is for the robot to help you with your exercises you'd rather have somebody cheerful that makes you want to do the exercises rather than very mechanical, I think it will encourage people to do more"*

Participants also suggested that the robot was just doing the same as a human equivalent would, in some cases making a interesting parallel regarding the potential for deception in those interactions too.

[G23]: *"I know it's been programmed to, and it kind of will ask that to everyone, but then you know I know from [therapy that therapists] do the same thing, they very much say hey how you doing regardless of whether they want to see you or not"*

[G18]: *"I knew [Pepper] didn't really care but when he said it it did kind of feel genuine, and it kind of made me feel like sometimes even when people ask, they don't really mean it or it's just to start a conversation"*

The small number of participants (3/52) who found the similarity or goodwill behaviours unacceptable suggested they felt the behaviours were disingenuous and unnecessary, even for the proposed application. However, on reflection they also commented how that might be a personal preference and they could imagine it might be a benefit for others:

[S13]: *"That pretending to interact with me...it just felt like a waste of time... [but] other people might feel like it that bit of social interaction, might be helpful for people who are on their own all day"*

(RQ1, RQ2) In Study 2, the overwhelming majority of participants elected that *none* of the robots they saw were deceptive, as can be seen in Figure 3.

5.2 Impact of Behaviour (Re-)Design

(RQ2) Here we present the results from Study 2 concerning variations in how the *Higher Risk*, *Lower Risk* and *control* dialogue strategies were perceived by participants. Repeated measures ANOVA analysis was used to check for significant differences across conditions, followed by Bonferroni post-hoc tests to compare individual

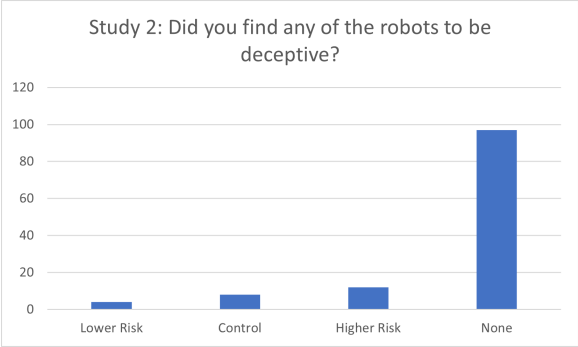

Figure 3: Frequency count of responses to the deception question in Study 2.

Table 3: Mean and standard deviation for all Likert-scale based measures across the three experimental conditions of Study 2. Scale abbreviations are taken from following text.

	Control	Lower Risk	Higher Risk
E	M = 3.66 SD = 0.81	M = 3.91 SD = 0.79	M = 3.91 SD = 0.85
G	M = 3.06 SD = 0.95	M = 3.72 SD = 0.94	M = 3.90 SD = 0.94
T	M = 3.52 SD = 0.70	M = 3.77 SD = 0.70	M = 3.80 SD = 0.77
L	M = 3.31 SD = 0.91	M = 3.87 SD = 0.84	M = 4.02 SD = 0.86
TM	M = 3.66 SD = 1.34	M = 3.78 SD = 1.20	M = 3.65 SD = 1.25
TA	M = 3.89 SD = 1.24	M = 3.94 SD = 1.12	M = 3.86 SD = 1.17
RM	M = 2.34 SD = 1.39	M = 2.71 SD = 1.39	M = 2.94 SD = 1.51
RA	M = 2.45 SD = 1.38	M = 2.88 SD = 1.42	M = 2.94 SD = 1.51

conditions in those cases. Mean and standard deviation results for all measures are presented in Table 3. Significant differences are identified below with the corresponding p-value and effect size (partial eta squared).

5.2.1 Credibility.

- Expertise (E) $F(2, 121) = 8.49, p < .001$ with small effect size (0.066): Higher Risk > Control ($p = .004$); Lower Risk > Control ($p = .002$)
- Goodwill (G) $F(2, 121) = 52.5p < .001$ with moderate effect size (0.304): Higher Risk > Lower Risk ($p = .043$); Higher Risk > Control ($p < .001$); Lower Risk > Control ($p < .001$)
- Trustworthiness (T) $F(2, 121) = 13.6, p < .001$ with small effect size (0.102): Higher Risk > Control ($p < .001$); Lower Risk > Control ($p < .001$)

5.2.2 Likeability (L) $F(2, 121) = 47.8, p < .001$ with small effect size (0.285).

- Higher Risk > Lower Risk ($p = .010$); Higher Risk > Control ($p < .001$); Lower Risk > Control ($p < .001$)

5.2.3 Therapist & Robot Responsibility.

- Robot Responsibility for Monitoring Patient (RM) $F(1.906, 121) = 19.860, p < .001$ with small effect size (0.142): Higher Risk > Lower Risk($p = .029$); Higher Risk > Control ($p < .001$); Lower Risk > Control ($p = .001$)
- Robot Responsibility for Advising Patient (RA) $F(1.906, 121) = 19.860, p < .001$ with small effect size (0.180): Higher Risk > Control ($p < .001$); Lower Risk > Control ($p < .001$)

No significant difference was found on therapist responsibility for monitoring the patient (TM) $F(1.742, 121) = 1.125, p = .321$ or therapist responsibility for giving advice to the patient (TA) $F(1.835, 121) = .508, p = .602$.

5.2.4 Most Motivating & Work Preference. The *Higher Risk* robot was most commonly selected as the most motivating robot and the preferred robot to work with (67/120 and 60/121 respectively). However, this wasn't unanimous, with approximately one third of participants (41/121) instead choosing the *Lower Risk* robot as being the most motivating. In addition, a number of participants who identified either the *Higher* or *Lower* risk robot as being the most motivating then suggested they would actually prefer to work with the control. The results are shown in Figure 4.

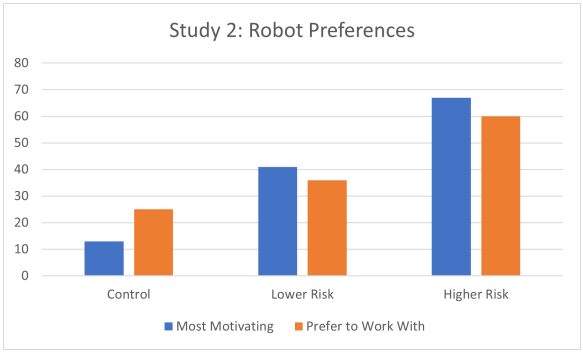

Figure 4: Count of participant choices for which robot was most motivating and which robot they would prefer to work with from Study 2.

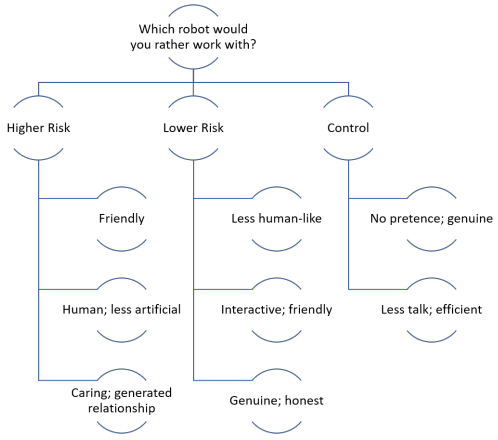

Figure 5: Emergent themes regarding participants' preferred choice of robot from Study 2.

Participants were asked to explain their choices using open-text comments. The resultant data were coded with the same Framework

method applied to Study 1 interview data. Figure 5 shows emergent themes regarding participants choice of which robot they'd rather work with. The reasons for selecting a particular robot seem to somewhat reflect the experimental manipulations e.g. with the *Higher Risk* robot being *more human-like*, the *Lower Risk* robot being more *honest/genuine* and the control having *no pretence* at all.

6 DISCUSSION

6.1 (RQ1) Assessing Ethical Risk in Typical SAR Behaviour

Here we use results from both studies to assess whether there is evidence of the ethical risks identified in Table 1, and, if so, what the likelihood and severity of those risks might be.

6.1.1 Deception. Participants in both studies overwhelmingly found the robot either *not deceptive* or *deceptive but acceptable*. The most common reason given was that participants *knew* it was *'only' a robot*, that robots don't *'have feelings'* and only do *what they are programmed to do*. Hence, participants felt they were not, and could not be, deceived about the robot's social or affective capabilities. Given that our participant pool specifically represents a *non-vulnerable* population, we cannot say the same would be true for all users; but these results at least suggest that the risk of deception is relatively low for the general population.

It is also interesting to note the slight differences in how Study 1 versus Study 2 participants answered this question. Specifically, Study 2 participants overwhelmingly suggested *none* of the robots they saw were deceptive, whereas a number of Study 1 participants found the robot to be *deceptive but acceptable*; even in the control condition. The *Higher Risk* condition of Study 2 was specifically designed to replicate those anthropomorphic (and potentially deceptive) behaviours showcased in Study 1. Therefore, one might expect that a number of Study 2 participants would identify at least the *Higher Risk* robot as being deceptive, in-line with the Study 1 results.

It could be that Study 1 participants, who interacted with the robot in person, were more aware of the potential for deception. Participant comments regarding how genuine the interaction 'felt' support this notion. For the Study 2 participants then, it might simply have been easier to dismiss the ethical risk of deception when watching the interaction rather than experiencing it first hand. This implies that in-person, exemplar interactions with a robot rather than online studies and/or e.g. abstract focus groups/user polling ought to be used for *user validation and verification* when assessing ethical risk.

Moving towards the potential for *over trusting*, a number of participants suggested it *was* deceptive to have the robot comment or give feedback on something it wasn't *actually* monitoring at high resolution (e.g. quality of movement). The small minority of participants who found the robot's social behaviour to be unacceptable specifically referred to this potential 'mismatch' between the robot's social interaction capabilities and its actual exercise monitoring capabilities. They suggested it was wrong (and unnecessary) to display these affective goodwill/similarity behaviours but then 'under-deliver' on the functional exercise monitoring.

6.1.2 Over Trusting. Qualitative data from Study 1 suggests participants were quite skeptical with regards to the robot's exercise monitoring capabilities, as they queried whether it could really give accurate performance feedback. Results on responsibility ascription to the robot from both studies further support this, with robot responsibility ascription typically being low, specifically much lower than the responsibility ascribed to the therapist. As such, the likelihood of *conscious* over-trusting, *specifically* in relation to the robot's social interaction capabilities, seems quite low.

6.1.3 Uncanny Valley. Arguably, of all the risks considered in this work, the *uncanny valley* might have both the highest likelihood and potential severity, in that it could result in users refusing to use the robot. Whilst we established that the risk of *deception* due to social behaviours is low, we also documented that a number of Study 1 participants *disliked* those behaviours, with a small minority even finding them actively unacceptable on that basis. However, the objective exercise repetition results (and much more positive feedback from other participants) in Study 1 strongly suggest that these behaviours fundamentally make the SAR more motivating and, hence, better at its intended function. This provides good motivation for RQ2 and Study 2, suggesting it is definitely worthwhile to explore whether such behaviours can be re-designed in a way which reduces this risk of uncanny valley effects without impacting on the robot's efficacy.

6.2 (RQ2) Impact of (Re-)Designing Robot Behaviours

Here we specifically reflect on the results from Study 2 to consider whether our attempt to re-design the SAR behaviours from Study 1 (i) actually resulted in lower ethical risk and (ii) impacted on the potential efficacy of the SAR as an exercise guide/motivational tool.

6.2.1 Reducing Ethical Risk. Given that the majority of Study 2 participants suggested they didn't find any of the robots presented to be deceptive, it is difficult to say whether our re-design of Study 1 behaviours (i.e. the *Lower Risk* robot condition) really reduced the risk of deception. However, qualitative data concerning participants' robot preferences does suggest that the *Lower Risk* robot was perceived to be less deceptive, with participants describing it as 'more honest' and 'genuine'. Similarly concerning the uncanny valley, the data also suggests participants perceived the *Lower Risk* robot as being less human-like than the *Higher Risk* condition. Regarding over-trusting, the *Lower Risk* robot was ascribed less responsibility for monitoring the patient than the *Higher Risk* robot, but more than the control robot, suggesting it may indeed offer a 'middle ground' between the two. Some responsibility ascription is obviously important if the robot is to be useful. As such, we would argue that our *Lower Risk* condition was a successful demonstration of how typical social (assistive) behaviours can be re-designed for reduced ethical risk.

6.2.2 Impact on Efficacy. In HHI, it is well established that credibility and likeability (measured as per our studies) directly correlate with how persuasive an agent is [6]. As such, we propose that our Study 2 results these measures, combined with participants' overall preferences for which robot was most motivating/which robot they

would rather work with offer proxy measurements for efficacy akin to our objective 'number of repetitions' measure in Study 1.

The *Higher* and *Lower Risk* robots were consistently rated as being more credible and likeable than the control, across all measures. This re-affirms the key result from Study 1; namely that social behaviour does positively impact on SAR efficacy, and is therefore an important and valuable design feature to have. The difference between these two conditions is less clear, with the *Higher Risk* robot scoring better on a subset of measures only, and with lower significant difference. As such, there is *some* suggestion that the *Higher Risk* behaviours *might* result in increased efficacy, but further work (ideally in the form of an in-person study) would be required to really test whether that resulted in any meaningful difference that justified the increased risk.

Participant choices for which robot was most motivating/which robot they would rather work with fails to add clarity on this. Whilst the *Higher Risk* robot was most commonly selected as both the most motivating and the preferred robot to work with, the results were certainly not unanimous. Further, qualitative data regarding these choices suggest it was specifically the experimental manipulations that informed these choices. Participants that preferred the *Higher Risk* robot specifically liked that it was more *human-like* and *caring*, whereas those who preferred the *Lower Risk* robot did so because it was *less human* and *more genuine and honest*. Further, the number of participants who said they would prefer to work with the *control* cannot be discounted, again giving reasons related specifically to its *lack* of social interaction.

In summary, the results suggest that attempting to reduce ethical risk in SAR behaviours likely will impact on efficacy. However, whether that impact is positive or negative is likely to vary across individual users based on their preferences for anthropomorphism and social robot interaction (or lack thereof).

6.3 (RQ3) Practical Implications for SAR Design and ERA

As discussed in Section 1, BS8611 identifies anthropomorphization and deception as (linked) ethical hazards, suggesting *unnecessary* anthropomorphism should be avoided. However, our results suggest that anthropomorphic robot behaviours are generally considered both non-deceptive and/or acceptable by the overwhelming majority of users. Further work is certainly required to carefully consider how the ethical risk posed by anthropomorphic deception might vary across different populations, but these results suggest a general consensus that anthropomorphic behaviours are not only acceptable but also perceived as being *important* to the overall function of a SAR.

BS8611 notes that anthropomorphism should be used *'only for well-defined, limited and socially-accepted purposes'*, but that in some cases it might be a necessary part of the functionality. At the very least our results demonstrate that SARs represent one of these socially-acceptable use cases. More broadly, they suggest that the main risk resulting from anthropomorphic behaviours (at least in a robot no more human-like than Pepper) is associated more with *user dislike* effects than with *deception*. Specifically, there are a number of users who might *prefer* to work with a robot that is more 'upfront' about its nature, a smaller number who'd rather have no unnecessary social interaction and a smaller minority again for whom the social behaviours are completely unacceptable. However, our results still suggest that actively anthropomorphic behaviour is (i) most preferred by the largest number of users and (ii) has the greatest positive impact on a SARs efficacy. As such, we propose a more sensible mitigation strategy for SARs would be to make anthropomorphism a customisable robot setting that can be tailored to the user's preference (something akin to the experimental conditions of Study 2).

7 CONCLUSION

In this work, we have applied the concept of *Ethical Risk Assessment* to an exemplar socially assistive robot, specifically considering risks associated with anthropomorphic robot behaviours as highlighted in a published standard for ethical robot design.

Based on a previous study in which we demonstrated the efficacy of anthropomorphic behaviour, plus a novel study designed exclusively for this work, we re-affirmed that anthropomorphic behaviours are crucial to the overall function of SARs. We also demonstrated that the actual ethical risk posed by such behaviours appears to be relatively low. We suggest *user dislike* effects actually represent the largest risk, as a minority of users are likely to find social robot behaviours so undesirable that they simply choose to avoid working with the robot. However, we found that attempting to reduce anthropomorphism (and hence ethical risk) would make the SAR more effective for some users, but less effective for others.

Overall, we suggest that SARs represent one use case for which the use of anthropomorphism and any related ethical risks are justified. However, as a practical strategy for reducing risk, we suggest designers should consider how to make the level of anthropomorphism a customisable trait that can be adapted to the user. We suggest that for SARs, this is a more appropriate mitigation strategy than simply avoiding anthropomorphism wherever possible. In future work, we hope to further analyse our detailed participant demographic data in order to explore whether user traits such as age, gender, health profile, experience with and attitude towards robots might be used to inform this recommended strategy of tailoring anthropomorphism to the user.

This initial work is ultimately intended to spark more research and discussion at this crossing point between practical SAR design and responsible, ethical robotics. As such, there are a number of limitations and related opportunities for future work. Firstly, our studies were all conducted with the Pepper robot, considering only changes in anthropomorphic dialogue rather than e.g. physical design. Participant acceptability of these behaviours in Pepper might not hold in robots with a more/less human likeness.

Secondly, our participant pools consisted only of *non-vulnerable* adults; and whilst they suggested these behaviours might be particularly acceptable for vulnerable populations, expert guidance is required to consider that further. Finally, given that Study 2 was conducted online, our results concerning the practical impact of re-designing SAR behaviours for reduced ethical risk are somewhat limited. A follow-up, in person study is required to demonstrate if/how the results we presented might translate into real world HRI behaviour.

REFERENCES

[1] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International journal of social robotics* 1, 1 (2009), 71–81.

[2] BSI. 2016. BS 8611:2016 Robots and Robotic Devices: Guide to the Ethical Design and Application of Robots and Robotic Systems.

[3] Mark Coeckelbergh. Fourth 2012. Are Emotional Robots Deceptive? *IEEE Transactions on Affective Computing* 3, 4 (Fourth 2012), 388–393. https://doi.org/10.1109/T-AFFC.2011.29

[4] David Feil-Seifer and Maja J Matarić. 2005. Defining Socially Assistive Robotics. In *Rehabilitation Robotics, 2005. ICORR 2005. 9th International Conference On*. IEEE, 465–468.

[5] Nicola K. Gale, Gemma Heath, Elaine Cameron, Sabina Rashid, and Sabi Redwood. 2013. Using the Framework Method for the Analysis of Qualitative Data in Multi-Disciplinary Health Research. *BMC Medical Research Methodology* 13 (Sept. 2013), 117. https://doi.org/10.1186/1471-2288-13-117

[6] Robert H. Gass and John S. Seiter. 2015. *Persuasion: Social Influence and Compliance Gaining*. Routledge.

[7] Rachel Gockley and Maja J Mataric. 2006. Encouraging Physical Therapy Compliance with a Hands-off Mobile Robot. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*. ACM, 150–155.

[8] Ryan Blake Jackson, Tom Williams, and Nicole Smith. 2020. Exploring the Role of Gender in Perceptions of Robotic Noncompliance. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20)*. Association for Computing Machinery, New York, NY, USA, 559–567. https://doi.org/10.1145/3319502.3374831

[9] Séverin Lemaignan, Alexis Jacq, Deanna Hood, Fernando Garcia, Ana Paiva, and Pierre Dillenbourg. 2016. Learning by Teaching a Robot: The Case of Handwriting. *IEEE Robotics Automation Magazine* 23, 2 (June 2016), 56–66. https://doi.org/10.1109/MRA.2016.2546700

[10] Norjasween Abdul Malik, Fazah Akhtar Hanapiah, Rabiatul Adawiah Abdul Rahman, and Hanafiah Yussof. 2016. Emergence of Socially Assistive Robotics in Rehabilitation for Children with Cerebral Palsy: A Review. *International Journal of Advanced Robotic Systems* 13, 3 (June 2016), 135. https://doi.org/10.5772/64163

[11] James C. McCroskey and Jason J. Teven. 1999. Goodwill: A Reexamination of the Construct and Its Measurement. *Communication Monographs* 66, 1 (March 1999), 90–103. https://doi.org/10.1080/03637759909376464

[12] James C. McCroskey and Thomas J. Young. 1981. Ethos and Credibility: The Construct and Its Measurement after Three Decades. *Central States Speech Journal* 32, 1 (March 1981), 24–34. https://doi.org/10.1080/10510978109368075

[13] Hannah R. M. Pelikan, Mathias Broth, and Leelo Keevallik. 2020. "Are You Sad, Cozmo?": How Humans Make Sense of a Home Robot's Emotion Displays. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20)*. Association for Computing Machinery, New York, NY, USA, 461–470. https://doi.org/10.1145/3319502.3374814

[14] Emmanuel Senft, Séverin Lemaignan, Paul E. Baxter, Madeleine Bartlett, and Tony Belpaeme. 2019. Teaching Robots Social Autonomy from in Situ Human Guidance. *Science Robotics* 4, 35 (Oct. 2019). https://doi.org/10.1126/scirobotics.aat1186

[15] Robert Sparrow. 2002. The March of the Robot Dogs. *Ethics and Information Technology* 4, 4 (Dec. 2002), 305–318. https://doi.org/10.1023/A:1021386708994

[16] L. Sussenbach, N. Riether, S. Schneider, I. Berger, F. Kummert, I. Lutkebohle, and K. Pitsch. 2014. A Robot as Fitness Companion: Towards an Interactive Action-Based Motivation Model. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. 286–293. https://doi.org/10.1109/ROMAN.2014.6926267

[17] Tijs Vandemeulebroucke, Bernadette Dierckx de Casterlé, and Chris Gastmans. 2018. The Use of Care Robots in Aged Care: A Systematic Review of Argument-Based Ethics Literature. *Archives of Gerontology and Geriatrics* 74 (Jan. 2018), 15–25. https://doi.org/10.1016/j.archger.2017.08.014

[18] Alan F. T. Winfield and Katie Winkle. 2020. RoboTed: A Case Study in Ethical Risk Assessment. *2020 5th International Conference on Robot Ethics and Standards (ICRES)* (Sept. 2020). arXiv:2007.15864

[19] K. Winkle, S. Lemaignan, P. Caleb-Solly, U. Leonards, A. Turton, and P. Bremner. 2019. Effective Persuasion Strategies for Socially Assistive Robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 277–285. https://doi.org/10.1109/HRI.2019.8673313