A Meta-analysis of Vulnerability and Trust in Human-Robot Interaction

PETER E. MCKENNA, Heriot-Watt University, UK
MUNEEB I. AHMAD, Swansea University, UK
TAFADZWA MAISVA, Heriot-Watt University, UK
BIRTHE NESSET, Heriot-Watt University, UK
KATRIN LOHAN, University of Applied Sciences of Eastern Switzerland, Siwtzerland
HELEN HASTIE, Heriot-Watt University, UK

In human-robot interaction studies, trust is often defined as a process whereby a trustor makes themselves *vulnerable* to a trustee. The role of vulnerability however is often overlooked in this process but could play an important role in the gaining and maintenance of trust between users and robots. To better understand how vulnerability affects human-robot trust, we first reviewed the literature to create a conceptual model of vulnerability with four vulnerability categories. We then performed a meta-analysis, first to check the overall contribution of the variables included on trust. The results showed that overall, the variables investigated in our sample of studies have a positive impact on trust. We then conducted two multilevel moderator analysis to assess the effect of vulnerability on trust, including: 1) An intercept model that considers the relationship between our vulnerability categories; and 2) A non-intercept model that treats each vulnerability category as an independent predictor. Only model 2 was significant, suggesting that to build trust effectively, research should focus on improving robot performance in situations where the users is unsure how reliable the robot will be. As our vulnerability variable is derived from studies of human-robot interaction and human-human studies of risk, we relate our findings to these domains and make suggestions for future research avenues.

CCS Concepts: • **Human-centered computing → Empirical studies in interaction design**; **HCI theory, concepts and models**; **Laboratory experiments**; **Field studies**; **User studies**;

Additional Key Words and Phrases: vulnerability, trust, risk, human-robot interaction

1 INTRODUCTION

Robots already play a big part in human society. They autonomously cut our lawns, assist surgeons during surgery [57], and provide travel information at airports [21]. Indeed, a recent analysis from Oxford Economics indicated that robots will contribute significantly to economic productivity, with a projected 20 million robots to be deployed in this sector by 2030 [7]. As robots become more sophisticated and ubiquitous, they will need to operate effectively alongside people. This is why considerable research effort has been put into examining how to improve people's trust in robots (e.g., [15]). With greater trust in robots comes a greater willingness from people

Authors' addresses: Peter E. McKenna, p.mckenna@hw.ac.uk, Heriot-Watt University, Edinburgh, UK, EH14 4AS; Muneeb I. Ahmad, m.i.ahmad@swansea.ac.uk, Swansea University, Swansea, UK, SA2 8PP; Tafadzwa Maisva, tafadzwa.maisva@hw.ac.uk, Heriot-Watt University, Edinburgh, UK, EH14 4AS; Birthe Nesset, bn25@hw.ac.uk, Heriot-Watt University, Edinburgh, UK, EH14 4AS; Katrin Lohan, katrin.lohan@ost.ch, University of Applied Sciences of Eastern Switzerland, Rapperswil, Siwtzerland; Helen Hastie, h.hastie@hw.ac.uk, Heriot-Watt University, Edinburgh, UK, EH14 4AS.

ACM 2573-9522/2024/4-ART
https://doi.org/10.1145/3658897

to share a workspace with their robot assistants, to collaborate with them, which will lead to better outcomes like increased productivity and team effectiveness. So, understanding the factors that contribute to successful and harmonious human-robot trust is useful for robot developers and industries seeking robotic solutions. Indeed, poorly calibrated trust can lead to over-trust or under-trust in robots can put users in danger, or lead to the rejection of robotic systems (see [23] for a review).

In this review and meta-analysis, we focus on a factor of trust that is often overlooked in the human-robot interaction (HRI) research literature, the *trustor's vulnerability*. Trustor vulnerability refers specifically to the vulnerability that the person giving the trust (i.e., the user or participants in an HRI experiment) feels, which we will refer to henceforth as "vulnerability" for simplicity. Vulnerability is considered a key element of trust and has often been referred to in studies of HRI without direct investigation. Our goal was to study the contribution of vulnerability to human-robot trust (HRT).

In reviewing the literature of vulnerability, we decided to develop a conceptual model to encapsulate the related concepts and themes. The model includes different categories of robot interactions, which in turn map onto different types of risk. Following the development of our conceptual model, we used a moderator meta-analysis to demonstrate the contribution of our vulnerability model against measured trust outcomes from just under a decade (January 2011- July 2020) of HRI trust literature. We take this approach as related meta-analysis have successfully summarised and generated useful insights from HRT research ([14, 15]).

Our results showed that HRI studies of robot performance (our first vulnerability category) were the most likely to lead to increases in trust. This suggests that effective human-robot collaboration is dependent on the reliability of the system. However, we also note that more work is required to examine the contribution of the other vulnerability categories included in our model.

In the following review and meta-analysis we:

- Make a theoretical contribution, by examining the role of vulnerability to human-robot trust.
- Using the literature, create a conceptual model for vulnerability in the context of HRI.
- Run a multilevel moderator meta-analysis, that accounts for the complex dependency structure and heterogeneity of the HRI studies identified for analysis.
- Compute: 1) A global analysis of the study effect sizes; 2) A moderator analysis with an intercept term that examined the relationship between vulnerability categories; and 3) A second moderator without an intercept term that considered each category of vulnerability as an independent predictor of HRT.
- Demonstrate that experiments in which the concept variable was the robot's performance (vulnerability category 1) were most likely to lead to trust improvements, thereby indicating the importance of robot reliability to trust development.

2 BACKGROUND

In the field of HRI, "trust" is often defined in terms of the relationship between the trustor (the person placing their trust in an other) and the trustee (the person fulfilling the trust request). Take the definition offered by Lee and See [29], where trust is understood as, "an attitude that an agent (robot or another person) will help achieve an individual's goals in a situation characterised by uncertainty and *vulnerability*". Similarly, trust has also been defined as "the willingness of a party to be *vulnerable* to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" [38]. Linking these two oft cited definitions is the theme of *trustor vulnerability*, suggesting that trust is not possible without a person putting themselves at the mercy of someone else.

However, there may be something unique about human-robot trust relative to human-human trust that makes the use of these definitions problematic. For example, relative to human colleagues, large robots are considered to pose more of a physical threat [46]. The trust that humans give to robots working alongside them on an

assembly line therefore might be qualitatively different to the type of trust given to other human colleagues nearby. Similarly, robot's operating as airport information kiosks may be supplied potentially confidential information by a traveller; how can the traveller be certain that the disclosed information is going to handled securely?

In the situations described above, the human user or operator makes a conscious decision whether to trust the robot, despite the uncertainty or ambivalence they might experience. Therefore, taking a closer a look at how vulnerability has been defined and conceptualised will serve as an initial guide about the features of the interaction between human trustor and robot trustee that are worth further attention.

According to the Cambridge Online Dictionary, vulnerability can be defined as, "able to be easily hurt, influenced [sic], or attacked" [5]. So, broadly, vulnerability refers to a person's susceptibility to physical damage, as well as how psychologically malleable or impressionable they are. Taking a philosophical approach, Mackenzie [34] conceptualises vulnerability as "ambivalent potential" because, as humans, we are never quite sure whether the outcome of our vulnerability is going to be positive or negative. If positive, vulnerability can lead to personal growth, like learning a new skill. When vulnerability is negative, it can also lead to increased risk and danger. This interpretation of ambivalence, as a cognitive conflict between negative and positive outcomes, is echoed in studies of HRI. Seemingly, greater levels of robot autonomy are a source of ambivalence, whereby individuals report mixed feelings about the benefits and costs of greater robotic automation [54]. So, vulnerability can be understood in terms of the conflict between positive and negative outcomes, with people and robots.

In works studying trust, vulnerability has been understood specifically in terms of the dynamic between the trustor and the trustee. In Meyer et al. [39], *positive vulnerability* refers to the trustor or trustee's willingness to admit mistakes, such as forgetting to file a document away for a colleague and then confessing to your error later. *Negative vulnerability*, on the other hand, refers to a person's protective or defensive behaviour in response to threat. An example would be a boss's overly emotional response to rumours about staff redundancies [39]. So, vulnerability may also relate to a persons' willingness to let their guard down, or in other words, their willingness to be vulnerable. Taking a slightly different angle, Nienaber et al. [42] focus on situational factors, defining vulnerability as either *active* or *passive*. Active vulnerability refers to a situation where a trustor deliberately discloses private or sensitive information to the trustee, whereas passive vulnerability refers to situations where the trustor relies on the trustee to perform a certain task. Together, these conceptualisations of vulnerability suggest it is rooted in a person's psychological assessment of others, the uncertainty that follows this assessment (e.g., ambivalent potential), and the setting of the interaction.

The literature of HRI and vulnerability paints a similar picture. A review of HRT by Lewis et al. [30] (p.151) state that, "…vulnerability is dependent on whether the interaction is: a) reliant on the performance of the robot in some element of the task (e.g., retrieving an item the participant has moved to complete the task); or b) where the robot makes a deliberate attempt to get the user/participant to do something (e.g., disclose a secret)". So, having reviewed multiple HRI studies of trust, these authors surmised that vulnerability related specifically to the robots ability to perform a task properly and the robot's attempts to solicit information or action from the human trustor. Law and Scheutz [28] also refer to the robot's reliability and information solicitation, but add, "…This uncertainty can leave people vulnerable; for example…interacting with large and heavy robots may cause a person to be physically vulnerable" (p.29). As well as alluding directly to trustor uncertainty in situations mentioned by Lewis et al. [30], Law and Scheutz [28] specifically mention the physical risks associated with large robot interactions. Together, these reviews indicate that vulnerability in HRI can be understood in terms of three factors: 1) the robot's task efficiency (or reliability); 2) whether the robot attempts to get the user to disclose information; and 3) the size and potential threat the robot poses. While this information is useful for indicating HRI specific vulnerabilies, there may be clues from other literary sources as to why they emerge as key features of vulnerability.

Studies of perceived threat in HRI, for example, offers a lens to better understand why users might feel vulnerable whilst engaged in a task with a robot. Broadly, these studies refer to related user and robot factors.

With respect to users, a recent study found that the level of threat people feel towards robots is predicted by individual differences in mindset flexibly: individuals with a more flexible mindset towards robots were more likely to engage in a future robot interactions, comparatively to those with a "fixed" mindset [2]. Characteristics of the robot also play a part, with autonomous robots considered more threatening and less trustworthy compared to non-autonomous robots, though this is also affected by the appropriateness of the deployment setting [65]. Further, robot appearance is also a key factor to perceived threat. Human-like robot forms (e.g., androids) are perceived to be more of a threat to people's jobs [63], and to people's sense of self and identity [8], to less human-like robot forms. Large robots are also considered to pose more of a physical threat to other smaller robot forms [46]. Thus, perceived threat impacts users evaluations of robots, which in turn will impact their sense of vulnerability. Assessing threat requires an assessment of the trustee, their intentions, and the setting - much like the vulnerability factors suggested by Law and Scheutz [28].

Moreover, both review articles reflect that self-disclosure (i.e., revealing personal or private information to the robot) is source of vulnerability in HRI. On this topic, there has been a considerable research effort studying the different aspects that might lead a person to reveal information to a robot that they might not otherwise. Robots with expressive faces have shown to be effective at encouraging elderly people to revisit sensitive topics in conversation [43]. In a study inducing negative mood using a video of the Wenchuan Sichuan earthquake in China 2008, researchers found that participants negative affect could be attenuated effectively after disclosure of their emotions to a social robot, more so than documenting their feelings in a journal [6]. Similarly, an robotic agent designed to provide conversation-based stress therapy was able to create a setting suitable for self-disclosure [1]. The type of information solicited by the robot also has an impact on people's willingness to share information. Barfield [3] showed that when the content for disclosure is potentially embarrassing, participants are more willing to disclose this information to other people than a robot agent. In all, these studies demonstrate that robot's are effective tools for soliciting information from people, and thus, self-disclosure with a robot a key type of user vulnerability.

Underpinning much of what has been discussed, is that interacting with a robot involves a degree of risk. Risk has been defined as, "any consciously or non-consciously controlled behaviour with a perceived uncertainty about its outcome" [56]. Similarly to Mackenzie [34], risk relates to situational uncertainty, the resolution of which based on subjective experience. Tying things together, it is possible that this uncertainty stems from the users assessment of the robot, based on nature of the interaction, how the robot acts, and the setting of the interaction. It is not surprising, therefore, that some HRI researchers have redefined trust by replacing the term vulnerability with risk [60].

At a higher level, risk has been described as domain specific, relating to either situational or relational features; the former refers to perception based on contextual factors, and the latter based on experience [55]. As our vulnerability literature led us to consider the user's evaluation of the trustee and the setting as well as the unique qualities of the trustee, we focused on *situational risk* definitions and domains. Of the domains suggested by Stuck et al. [55] we selected those that were evident from the studies identified in our literature review: *performance risk*, *privacy risk*, *financial risk*, *physical risk*, and *security risk*.

- Performance risk: The assessment that task engagement can have negative consequences.
- Privacy risk: The assessment that an activity may compromise the personal information of an individual.
- Financial risk: The assessment that a situation may lead to monetary loss.
- Physical risk: Situations that are judged to be dangerous to a person's physical health.
- Security risk: The evaluation that an activity could be susceptible to criminal interference.

Altogether, our journey to discover what vulnerability means in the context of HRT led us to consider the role perceived threat, self-disclosure, and risk in greater detail. Rather than proposing a single usable definition of vulnerability, implemented the knowledge gained in our literature review to create a set of vulnerability

categories that encapsulated the key themes. Below we describe these vulnerability categories and provide a graphic for clarity.

2.1 Creation of vulnerability categories

To establish our vulnerability categories, we included the scenarios suggested by Lewis et al. [30] and Stuck et al. [55]. We were also led by our literature search and review, showing that studies of financial games (e.g., the Trust Game) were well represented in the literature base, and mapped onto the *financial* risk category [55]. Risk was also a prominent feature to HRI studies of vulnerability, we cross-referenced (through experimenter agreement) the risk categories offered by Stuck et al. [55] onto our vulnerability categories. Our resulting vulnerability categories are as follows:

(1) When the user must rely on the robot to complete a task and are unsure how competent the robot is.
(2) When the robot asks for sensitive or personal information (not financial) from the user and the user is unsure how that information is going to be used.
(3) Economic games with a degree of financial risk.
(4) When the user interacts with a large, heavy robot in proximity.

We felt that these categories captured the essence of vulnerability in HRI. As well as being informed by large literature reviews, that also capture other related factors that might affect research participants decision making in HRI studies, like perceived threat and self-disclosure. Below we describe the method we adopted to test the predictive quality of our vulnerability categories, as well as detail each study selected for the analysis, and the process of conducting a moderator meta-analysis.

3 METHOD

Here, we describe the process we followed to conduct the moderator meta-analysis on the literature.

3.1 Collating the bibliography

We followed a similar procedure to Hancock et al. [15], as their meta-analysis provided a roadmap for navigating the HRI and trust literature. Our database search included the terms "*human-robot interaction*" AND "*robot*" AND "*trust*". The number of records returned were as follows: ScienceDirect (N= 2,974); ACM Digital Library (N = 2,159); IEEE (N = 204); APAPsychInfo (N = 58); APAPsychArticle (N = 18); Applied Science and Technology (N = 16). This generated a total of 5,309 studies. Publication date was set between Jan 2011 - June 2020 as our intention was to follow up on the meta-analysis of HRT conducted by Hancock et al. [15], which examined all relevant papers up to the year 2010. We also extended our publication timeframe to June 2020 (rather than January 2020), to capture articles published in the summer robotics and AI conference run. Study duplicates were removed using the duplicate omission tools in Mendeley[1], and then again using JabRef[2].

3.2 Data screening

Initial screening of abstracts was conducted online using the web tool Rayyan QCRI [44]. Rayyan QCRI is designed for managing and organising large corpus of studies for literature screening. In Rayyan QCRI's user interface, article abstracts are identified from the .bib file and enlarged for easy reading. All 5,309 article abstracts were screened by the first author, who used Rayyan QCRI's search tool to inspect for; a) the terms "*robot*", or "*agent*", or "*automation*" or the use of "*trust*" or "*trustworthiness*" as a dependent variable.

Records that were not empirical (i.e., book chapters, conference workshop invitations) or irrelevant (i.e., unrelated topics) were excluded (N = 4,726). Study abstracts that did not clearly state the nature of the interaction,

[1]https://www.mendeley.com
[2]https://www.jabref.org

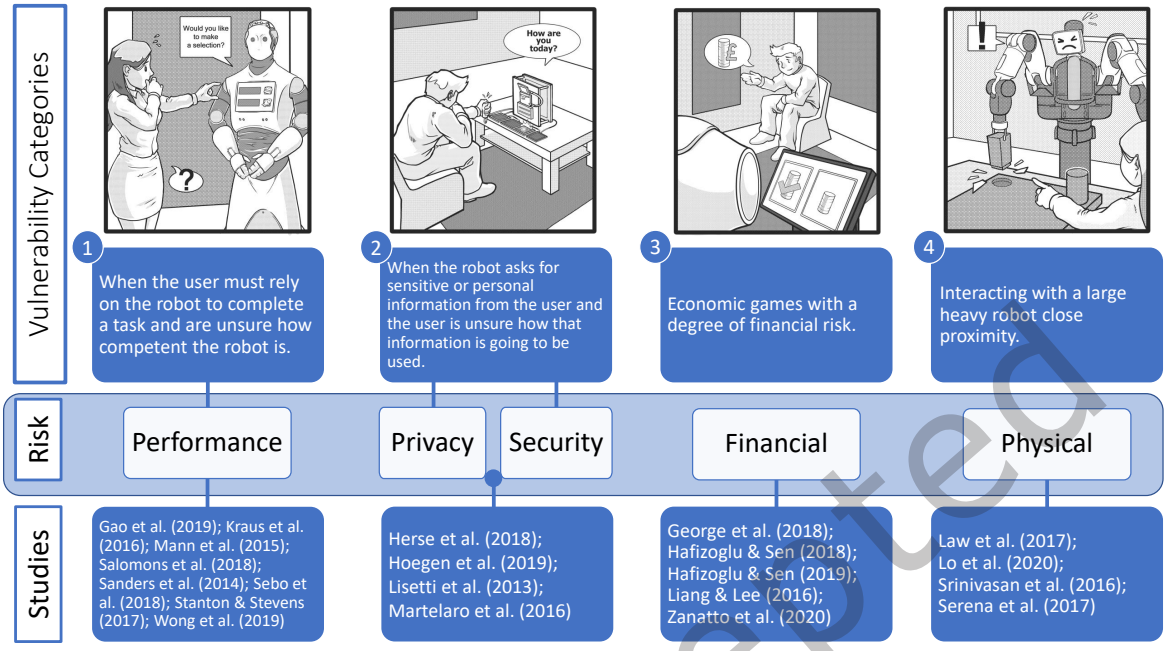

Fig. 1. Conceptual Mapping Between Vulnerability and Risk. Vulnerability categories inspired by the reflecitons from Lewis et al. [30] and Law and Scheutz [28], whilst risk categories were selected from Stuck et al. [55] and cross-referenced against the vulnerability categories through researcher agreement.

the dependent variable(s) or either were marked for further screening. A final round of detailed screening involved searching the article PDF for the term "*trust*". This initial screening process reduced the returned records from 5,313 to 583 (10.97%). At phase two of screening, the first and second authors (both active researchers in the field of HRI) independently screened the remaining 583 studies. Prior to this screening phase, it was agreed that inclusion decisions would be based upon assessment of the Abstract and Method sections relevance, and by searching the document for the term "*trust*". Once each author had finished marking each study for inclusion, a moderation meeting was held, where inclusion conflicts were resolved through discussion. This process resulted in a final list of 85 articles to be considered for the meta-analysis.

To finish, authors 1 to 4 were assigned a batch (roughly 20 each) of articles from the final list of 85 for data and study characteristic extraction. With respect to data extraction, studies had to report: 1) the subgroup sample sizes (sample size per condition); 2) subgroup means; and 3) subgroup standard deviations. Articles that did not contain enough necessary data to be included in the study were omitted[3], bringing the final number of included studies to 21 (see Figure 2 for the PRISMA flow diagram of study screening). In terms of study characteristics, we logged information related to the source of the dependent variable and study design (see Table 1). Other characteristics collected but not included in the present work included robot platform, questionnaire/survey name, with a view to be used in future related work.

Of the included 21 studies, we calculated 36 effect sizes. In terms of study design, 15 studies were between subjects (71.4%), 3 were within-subjects (14.3%), and 3 were mixed-design (14.3%). With respect to their dependent

[3]See Section "Studies Providing Incomplete Data" below for more details.

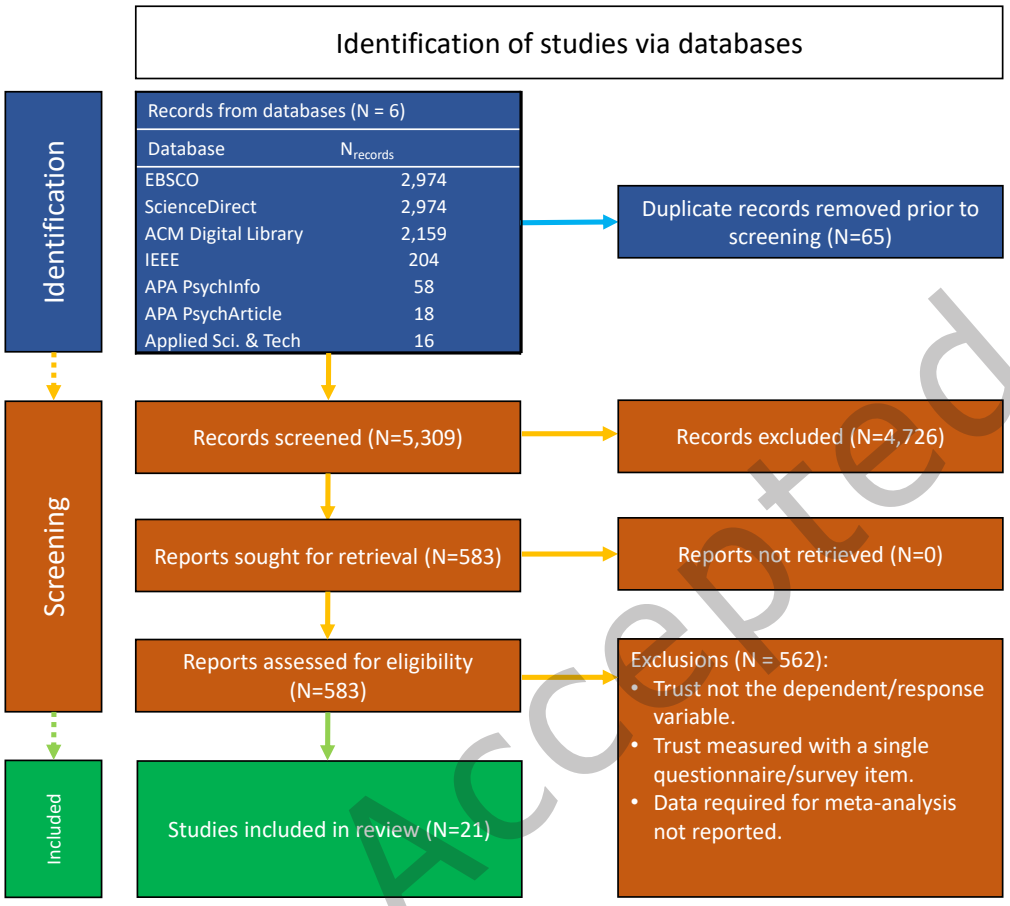

Fig. 2. PRISMA Diagram of Literature Review Process. *Author 1 screened the articles for the analysis.

variables, 3 studies (14.3%) measured trust only through objective behavioural measures, 9 studies (42.9%) used a combination of question items and objective behavioural measurement, and 9 studies (42.9%) used question items only.

3.3 Vulnerability Categorisation

Our vulnerability categories, their related risk types, and the associated studies are shown in Figure 1. A summary of the selected studies and the data extracted is provided in Table 1. To assign categories, authors 1 and 2 independently assigned categories to each of the 21 studies. When both authors' assigned categories matched, these values were used as the final vulnerability category for the study. For studies where there was a disagreement between the authors, a moderation meeting was held to discuss each case and to reach an agreement. Often these disagreements arose because of the broadness of category 1 ('When the user must rely on the robot to complete a task and are unsure how competent the robot is'), as there was an element of variable robot performance in each of the other categories. Thus, category assignment agreement required a nuanced appreciation of the study characteristics (e.g., if it included a financial incentive), using the vulnerability categories chosen for the

meta-analysis. For instance, in Lo et al. [33], the performance of the robot varies depending on which navigation algorithm condition the participants are assigned to, but in the context of the vulnerability categories it was better suited to category 4 ('When the user interacts with a large, heavy robot in proximity').

3.3.1 Category 1: When the user must rely on the robot to complete a task and are unsure how competent the robot is. Of the studies included in the meta-analysis, seven (33.33%) were included in category 1, and included studies of cooperation, shared control, and robot persuasion. Studies of cooperation included an investigation of user's perception of a small robot fitness instructor [36] and a robot railroad game assistant [49]. Experiments studying shared control between the user and agent control of a drone to complete a set of objectives [62], reaching a destination in a driving simulation with the option of switching to automated driving [24], and shared control of a drone in a simulated emergency evacuation [48]. This category also included experiments where a robot influenced participants' decision-making, including a card matching game where three robots offered answers that varied in agreement [47], and a task where the robot's gaze behaviour was designed to exert pressure on participants' final decision in a game of chance [53].

3.3.2 Category 2: When the robot asks for sensitive or personal information from the user and the user is unsure how that information is going to be used. Four studies of the 21 (19.04%) examined how requesting sensitive or personal information from the user affected trust. Often these requests for information were interleaved into a task to build trust prior to making the request. This was the case for a study examining human-robot collaborative LED circuit building [37] and a task where participants planned future events with an agent [20]. This category also included a study of alcohol consumption therapy agent, where participants shared personal information about their drinking habits [32]. In another study, the robot attempted to elicit restaurant preferences from participants [19]. Lastly, we included a study of participants decision to let a robot in or out of a dormitory [50], as the robot made a specific request to enter or exit a restricted area.

3.3.3 Category 3: Economic games with a degree of financial risk. Five of the 21 (23.8%) studies came under the category of economics games involving financial risk. To examine trust, these studies manipulate the sequence of events in the game, the trustworthy behaviour of the trustee, and the agent/robot presentation. In their study of trust, Liang and Lee [31] modified the level of risk involved in a human-robot collaborative chance based card game. In another, robot payouts to participants in the investment game were either fair or unfair [64]. Some experimenters examined whether priming the reputation of the robot (as either fair or unfair) affected participants investments in a trust game [12, 13], whilst others examined whether participants displayed the same type of fairness assessments towards a robot in a virtual environment [11].

3.3.4 Category 4: Interacting with a large, heavy robot in close proximity. Included in the large robot interaction category were three studies (14.3%). Studies in this category included collaborative tasks where the robot and human worked together to reach a destination [33], or where a home assistance robot completed a series of tasks for the participant [52], and where the robot selected items for recycling [27].

3.4 Vulnerability category summary

As a first step to examining the impact of vulnerability in HRI we imputed these categories as a variable in our meta-analysis of HRT studies. Because little is known about the independence or dependence of these categories and how risk inter-plays with each, our analysis considered the relationship between and within these categories.

Table 1. Summary Statistics of Meta-analysis Studies. The summary includes the study authors, sample size (control and treatment group per effect size), the effect size variable, the assigned vulnerability category, the source of the dependent variable, and the Hedges *g* effect sizes. Vulnerability categories: 1 = When the user must rely on the robot to complete a task and are unsure how competent the robot is; 2 = When the robot asks for sensitive or personal information from the user and the user is unsure how that information is going to be used; 3 = Economic games with a degree of financial risk; 4 = Interacting with a large, heavy robot in close proximity.

Author	Design	Sample Size	Effect Size Variable	Vulnerability Category	DV Source	Hedges *g*
Gao et al. [10]	mixed-design	24	Algorithm with learning vs algorithm without learning	1	survey/questionnaire items	3.028
George et al. [11]	within-subjects	42	Human like avatar vs robot	3	survey/questionnaire items	-0.142
Hafızoğlu and Sen [12]	between-subjects	100	Negative reputation vs positive reputation	3	behavioural; survey/questionnaire items	0.502
Hafızoğlu and Sen [13]	between-subjects	200	Negative vs positive past experience with robot	3	behavioural; survey/questionnaire items	0.322
Hafızoğlu and Sen [13]	between-subjects	115	No past experience vs positive past experience; 3rd interaction	3	behavioural; survey/questionnaire items	0.377
Hafızoğlu and Sen [13]	between-subjects	115	No past experience vs positive past experience; 5th interaction	3	behavioural; survey/questionnaire items	0.517
Herse et al. [19]	between-subjects	48	Preference elicitation vs no preference elicitation	2	behavioural; survey/questionnaire items	0.113
Hoegen et al. [20]	between-subjects	14	No conversational matching vs high consideration	2	survey/questionnaire items	0.046
Hoegen et al. [20]	between-subjects	16	No conversational matching vs high involvement	2	survey/questionnaire items	-0.003
Stanton and Stevens [53]	mixed-design	25	Averted gaze vs constant gaze (females only)	1	behavioural	-0.546
Stanton and Stevens [53]	mixed-design	26	Averted gaze vs situational gaze (females only)	1	behavioural	0.143
Kraus et al. [24]	between-subjects	11	Voice only vs social NAO	1	survey/questionnaire items	1.339
Kraus et al. [24]	between-subjects	11	Voice only vs non-social NAO	1	survey/questionnaire items	0.295
Law et al. [27]	between-subjects	36	High surprise vs low surprise	4	behavioural; survey/questionnaire items	0.259

Continued on next page

Table 1 – continued from previous page

Author	Design	Sample Size	Effect Size Variable	Vulnerability Category	DV Source	Hedges *g*
Liang and Lee [31]	between-subjects	48	Descriptive information vs user-generated content (UGC)	3	behavioural	-0.280
Liang and Lee [31]	between-subjects	48	Descriptive information vs robot-generated content (RGC)	3	behavioural	-0.218
Lisetti et al. [32]	between-subjects	55	Textual vs non-empathic	2	survey/questionnaire items	-0.154
Lisetti et al. [32]	between-subjects	56	Textual vs empathic	2	survey/questionnaire items	0.657
Lo et al. [33]	within-subjects	26	Legible motion vs user-aware motion algorithm	4	survey/questionnaire items	0.115
Mann et al. [36]	between-subjects	65	Tablet vs robot	1	survey/questionnaire items	0.625
Martelaro et al. [37]	between-subjects	30	Low robot vulnerability vs high robot vulnerability	2	survey/questionnaire items	0.642
Martelaro et al. [37]	between-subjects	31	Low robot expressivity vs high robot expressivity	2	survey/questionnaire items	0.489
Salomons et al. [47]	between-subjects	30	Not know robot's preliminary answer vs knew robot's preliminary answer	1	behavioural; survey/questionnaire items	1.197
Sanders et al. [48]	mixed-design	49	Minimal v contextual	1	survey/questionnaire items	0.176
Sanders et al. [48]	mixed-design	49	Minimal v constant	1	survey/questionnaire items	0.394
Sanders et al. [48]	mixed-design	49	Contextual vs constant	1	survey/questionnaire items	0.390
Sebo et al. [49]	between-subjects	105	Neutral comments vs vulnerable comments	1	behavioural; survey/questionnaire items	0.808
Serena et al. [50]	between-subjects	19	Non-food robot vs food robot	4	behavioural	-0.163
Srinivasan et al. [52]	between-subjects	32	Multiple teleoperators vs autonomous robot	4	behavioural; survey/questionnaire items	-0.588
Srinivasan et al. [52]	between-subjects	32	Multiple teleoperators vs single teleoperator	4	behavioural; survey/questionnaire items	-0.985
Wong et al. [62]	within-subjects	14	Baseline voice vs male voice	1	behavioural; survey/questionnaire items	0.000

Continued on next page

Table 1 – continued from previous page

Author	Design	Sample Size	Effect Size Variable	Vulnerability Category	DV Source	Hedges *g*
Wong et al. [62]	within-subjects	12	Baseline voice vs female voice	1	behavioural; survey/questionnaire items	0.304
Zanatto et al. [64]	between-subjects	59	Unfair for both players vs unfair for human player	3	behavioural; survey/questionnaire items	0.646
Zanatto et al. [64]	between-subjects	60	Unfair for both players vs Unfair for robot player	3	behavioural; survey/questionnaire items	2.348
Zanatto et al. [64]	between-subjects	59	Unfair for both players vs unfair for human player	3	behavioural; survey/questionnaire items	0.901
Zanatto et al. [64]	between-subjects	60	Unfair for both players vs unfair for robot player	3	behavioural; survey/questionnaire items	0.030

3.5 Effect Size Calculation and Analysis Inclusion Criteria

We followed the guidelines given by Harrer et al. [17] for study inclusion. Hedge's *g* was calculated for both independent and dependent effect sizes. We opted for Hedge's *g* as it is an adjusted version of Cohen's *d* that accounts for potential bias due to unequal study sample sizes. As can be seen in Table 1, our effect size sample sizes varied greatly, with a minimum of 11 and maximum of 200. Furthermore, only two studies in the pooled literature ([20, 31]) used control group comparisons, and we were confident the use of multilevel meta-analysis would help to reduce the risk of unit-of-analysis error.

For within-subjects experiments that included non-correlated independent data (e.g., data from different groups), only the independent data were entered into analysis. For example, in Wong et al. [62] the experimenters analysed the contribution of same gender or different gender voice pairings between agent and participant performance. Gender is a quasi-experiment independent variable that generates independent data. So, we extracted data exclusive to each gender for effect size calculation, including the baseline comparison for each group against the condition of agent voice gender alignment; i.e., male participant's control condition (no agent voice) vs male participant's alignment with agent voice (male agent voice) and the effect on trust.

For studies with mixed designs, we considered whether the within-subjects variables were of theoretical interest. Variables that were not related to the robot behaviour (e.g., task complexity) for example were not considered in favour of comparisons between different robot behaviours. For instance, in Stanton and Stevens [53], we compared the means between groups across the whole experiment, omitting comparisons of the within-subject's variable, task difficulty. For Stanton and Stevens [53], the robot using averted gaze was treated as the control condition - as suggested by the authors - which was compared against the other two robot gaze conditions, constant and situational.

3.5.1 Studies Providing Incomplete Data. Some studies did not report all the required statistics for effect size generation. Among them included the work of Sanders et al. [48], whose subgroup sample size was not reported, meaning that SMD could not be calculated from the data provided. This is a common problem to meta-analysis, and various methods have been explored to overcome the loss of data [22]. In the case of Sanders et al. [48], we entered an estimate of the subgroup sample size value by dividing the total sample by the number of groups

Table 2. Comparison of Main Meta-analysis model and Model with Influential Cases Removed.

Analysis	***g***	**95%CI**	***p***	**95%PI**	I^2	**95%CI**
Main analysis	0.36	0.13—0.59	<0.05	-0.74—1.45	70.3	58.3—78.8
Infl. Cases Removed	0.33	0.19—0.46	<0.001	-0.12—0.78	32.4	0—56.2

(e.g., 73/3). We took this approach based on the work of Kambach et al. [22], who found, through meta-analysis simulations, that this replacement method is safe from incurring significant statistical alterations to analysis outcomes [22]. Statistical solutions to boost our study numbers were pursued where necessary.

There were also studies reporting non-parametric tests (e.g., [48]) due to small sample sizes. As these studies could not satisfy the criteria for standardised mean difference (SMD) calculation, they were also omitted.

4 RESULTS

The meta-analysis process followed the recommendations from Harrer et al. [17] for a random effects meta-analysis and multilevel moderator meta-analysis. A random effect model was chosen to account for different samples, experimental designs, and dependent variable choice. The multilevel moderator meta-analysis was computed to examine the moderating effect of vulnerability category on global effect size variance. We analysed vulnerability in two ways: 1) with an intercept term, to examine the relationship between our vulnerability categories; 2) without an intercept, to examine the contribution of each vulnerability category to HRT independently.

4.1 Influence and outlier analysis

Before proceeding with the meta-analysis, we conducted tests of study influence, outlier analysis, and respective plots were generated to determine the effects sizes to be included in the final analysis. Studies were considered outliers when the upper bound of 95% CI was lower than the lower bound of pooled effect CI (i.e., minuscule effects) and when the lower bound of the 95% CI was higher than the upper bound of the pooled effect confidence interval (i.e., extremely large effects). Four effect sizes were excluded based on these criteria: one from Zanatto et al. [64], one from Gao et al. [10], and two from Srinivasan et al. [52]. We report both the Main Meta-analysis Model (including outliers) and our Influential Cases Removed Meta-analysis model for comparison. Results from the analysis excluding influential cases is presented in the Forest plot (see Figure 3).

As you can see in Table 2, the I^2 value reduced from 70.3% to 32.4% following the removal of influential cases. I^2 represents the heterogeneity of effect sizes between studies in the sample, with lower I^2 values indicating lower effect size heterogeneity. To improve our meta-analysis inferences, we proceeded with a model that omitted influential cases and with a lower I^2 (though this practice is controversial in meta-analysis; see [59] for a discussion on influential and outlier case treatment in meta-analysis).

4.2 Multilevel Moderator Meta-analysis Model composition

Data were analysed in R [45] using the "metafor"[58] and "dmetar" [18] packages, following the analysis process outlined by Harrer et al. [17]. The models outlined in the analysis are three-level models to account for dependence introduced by study authors (e.g., collecting data from multiple sites, multiple comparisons to a control group, using different dependent variables to measure trust).

Our multilevel meta-analysis model was specified as follows:

- Hedges g was provided as the calculated effect size.
- We imputed the variance of the calculated effect sizes by taking the squared standard error of the effect size.

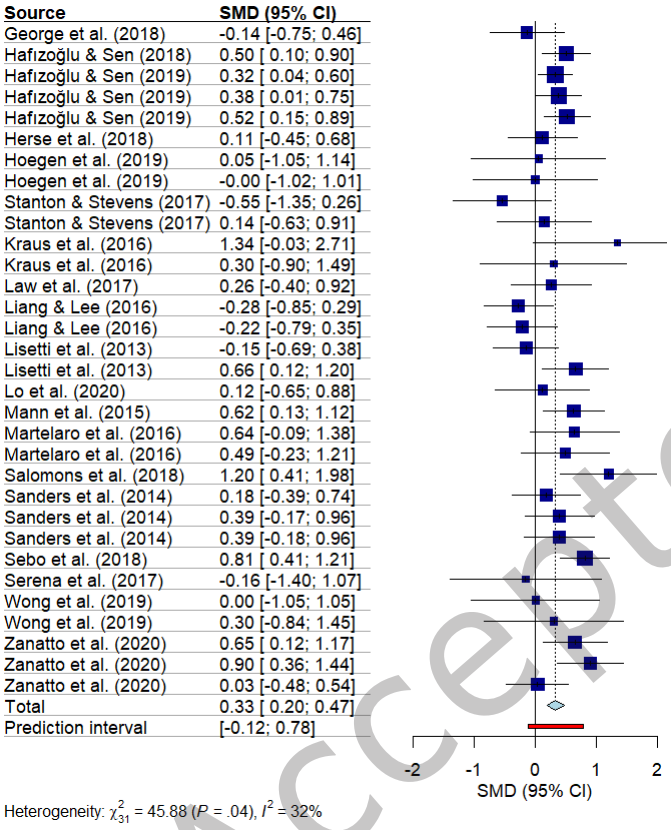

Fig. 3. A forest plot of the studies included after removal of influential cases. The plot details each study's effect size (Hedges g) and 95% CI.

- We imputed the study author name as the random effect (or random intercept). By doing so, the model assumed unique intercepts per study. In the model specification, we assume that individual effect sizes are nested within studies.
- Regression coefficients were tested using a method similar to Knapp-Hartung.
- The Restricted Maximum-Likelihood (REML) method was adopted to estimate the model parameters.

For the full multilevel meta-analysis script, see the file "data_analysis.R" in the project's OSF repository here.

We imputed a random intercept that considered the independence of both study author ("studlab") and each individual effect size in the analysis ("ef_num"). We used the restricted maximum likelihood (REML) model fitting procedure for meta-analysis, as recommended by [26].

To decide whether to include the nested individual effect sizes in the analysis, we compared the three-level or two-level models (or simple random effects model). This model comparison analysis indicated that the two-level model was more accurate. The two-level model was significant, showing that effect size Hedges g = 0.331 (95%CI: 0.19-0.47; p < 0.001).

To progress to the moderator analysis, we added our vulnerability variable as the model moderator. The results from our two-level meta-analysis modelling with moderator variable are shown in Table 3.

Table 3. Multilevel meta-analysis with vulnerability categories as a moderator variable. Coefficients represent the vulnerability categories: 1 = When the user must rely on the robot to complete a task and are unsure how competent the robot is; 2 = When the robot asks for sensitive or personal information from the user and the user is unsure how that information is going to be used; 3 = Economic games with a degree of financial risk; and 4 = Interacting with a large, heavy robot in close proximity.

			Moderator analysis model	
			Intercept Included	**Intercept Excluded**
Coefficient	**Studies**	**Effect sizes**	**Est. [SE]**	**Est. [SE]**
1	7	15	**0.49[0.16]***	**0.49[0.13]*****
2	4	7	0.26[0.19]	0.26[0.20]
3	5	10	0.27[0.22]	**0.26[0.14].**
4	3	3	0.13[0.19]	0.13[0.29]

Intercept Included model output: $QM(3) = 2.19, p = 0.53$.
Intercept Excluded model output: $\boldsymbol{QM}(4) = 18.60, \boldsymbol{p} < 0.001$.
*Significance codes: *** = p < 0.001; ** = p < 0.01; * = 0.05; . = p < 0.1.*

4.3 Moderator analysis of Vulnerability

To assess the contribution of our vulnerability category to HRT we conducted a random-effects multilevel moderator meta-analysis.

We ran a moderator meta-analysis, including an intercept and excluding an intercept. For moderator meta-analysis including the intercept, the intercept assumes that the average effect size is meaningful, even in the absence of the moderator. Excluding the intercept on the other hand assumes that the moderator has a direct effect on the effect size, and that there is no meaningful average effect without the moderator. In other words, the inclusion of an intercept assumes some kind of relationship between the moderator as a whole and the average effect size, whereas the model without an intercept assumes that the moderators independently modify the effect size, but not as a single grouped moderator.

We took this approach because vulnerability is not a well-defined concept, and it is possible that people react to the different scenarios included in our vulnerability categories uniquely, that cannot be captured through a single moderator variable. Thus, our analysis indicated whether to conceptualise vulnerability as a single construct, or as a set of independent categories.

Results from Model 1 (Intercept Included) indicate that the omnibus test of all levels of vulnerability category was non-significant, $QM(3) = 2.19, p = 0.53$, and therefore, that levels 1 to 4 of vulnerability category did not differ significantly to one another. Model 2 (Intercept Excluded) examines the contribution of each level of Vulnerability category as a predictor variable. This model was significant, $QM(4) = 18.60, p < 0.001$, thus, each level of vulnerability category independently has a significant effect on HRT (see 3 for analysis results).

5 DISCUSSION

This meta-analysis is the first of its kind to examine the contribution of vulnerability to trust in HRI. To do so, we defined four vulnerability categories based on related HRT review reflections [28, 30], that encapsulated the related concepts of perceived threat [8, 46, 65], self-disclosure [1, 3, 6, 43], and risk [55].

Prior to examining the predictive value of vulnerability, we ran a global meta-analysis of the studies identified from the screening process. We found a positive global effect of the independent variables investigated in studies of trust and HRI. A variety of variables were studied, including the robot's voice [24], gaze behaviour [53],

locomotion style [33], and interaction style [37] (for a full list of the HRI variables tested, see Table 1). That is, despite the range and variety of approaches taken by researchers, studies of HRT tend to indicate improvements in trust. However, it is also important to note that some effect sizes were negative [11, 20, 31, 32, 50, 52, 53], indicating that not all approaches resulted in improved trust as intended. Future work can refer to the effect sizes and study details provided here to make more informed choices about their robot or agent design.

The findings from multilevel moderator meta-analysis show that, when treated as a set of independent categories, studies of robot performance (i.e., vulnerability category 1) significantly predict trust. One way to interpret this finding is that the sense of vulnerability induced in the included studies was such that it did not adversely affect the user's trust rapport or intention to trust the robot. Taking a closer look at the variables investigated in vulnerability category 1 and relating it them to the concepts covered in the literature may serve to unpack this claim.

Vulnerability category included studies "when the user must rely on the robot to compete a task and are unsure how competent the robot is". So, vulnerability relates to the user's assessment of the robot's reliability in this category, and the uncertainty experienced in this process. There were a few studies that modified the robots behaviour, including a robot with a user aware locomotion algorithm [10], a robot with varying gaze movements during a game of chance[53], altering the robot's voice [24, 62], and varying the robot's dialogue during a collaborative task [48, 49]. These manipulations are geared toward enhancing user's experience, by designing and testing behaviour that is more familiar to user's; behaviours that are more socially attuned that could reduce the sense of uncertainty, and therefore, participants sense of vulnerability. As such, it is not surprising that the increased anthropomorphic features included in these experiment led to increases in trust [41].

Other studies in this category focused on the robot's form [36] and the user's prior knowledge [47]. It could be argued that the presence of a small exercise robot was considered a suitable deployment for a robot of that form, thereby leading to more positive appraisals without perceived threat [46]. In the second study, the manipulation of users prior knowledge perhaps poses the biggest threat to trust, as the sense of uncertainly could have adversely affected their vulnerability. What the result does indicate, is users' evaluations of a robot are more positive when their task is disambiguated with the supply of relevant information.

Generally, these HRI studies are designed to challenge users' expectations and confidence in the robot. A clue to why they were the most likely to lead to trust can be found by evaluating the experimental manipulations against the concepts and themes highlighted in our vulnerability review. Firstly, according to Nienaber et al. [42], these studies deal with passive vulnerability, whereby the trustor is beholden to the performance of the trustee. "Passive" is useful term here, as, in each study, there was no intention to manipulate the threat posed by the robot, or to solicit information from users. Perhaps it is because these studies focused mainly on the preformative features of the robot in the task that improvements in trust were found.

Taking this forward, perhaps the best way to develop trustworthy systems is to focus on performance related vulnerabilities. That is, user's trust in the robot is closely linked to their assessment of the robot's ability to complete a task. The study characteristics provided above suggest that this assessment of performance and the trust that follows can be affected by the robots behaviour and physical design. In light of this, stakeholders seeking effective robot solutions should focus on the preformative aspects of their system, and take inspiration from the different characteristics highlighted in the literature. For instance, robots operating in close quarters with humans would benefit from user aware navigational systems to build trust.

Moving on, the findings show an effect approaching significance for the contribution of vulnerability category 3, "Economic games with a degree of financial risk", to trust. It is possible that more research on the topic may eventually demonstrate statistical significance, so we discuss further. A significant result would indicate that people have a tendency to put their trust in a robot during a financial exchange. In these tasks, the robot's generosity is often manipulated, whereby it gives either a large or small payout to the human or robot player (e.g., see [64]). In western societies, trust in banks is considered a vital component to running an effective economy

[9]. So, it is conceivable that westerners view the sharing and competition for financial resources as culturally appropriate relative to more collectivist cultures [51]. Experiments with financial games and robots may represent an extension of this normalisation, represented as a tendency towards trusting others in financial transactions. Indeed, Malle and Ullman [35] indicate that trust can be understood in terms of competence and benevolence, whereby users willingness to take financial risks with a robot could be a result of their perceived benevolence of the robot. One way to unpack this further would be to conduct a cross-cultural HRI study between cultures with unique views on financial economics.

Taking the results at face-value would indicate that participants' vulnerability in economic games is not conducive to building trust with robots. Indeed, the inclusion of a monetary incentive can change the dynamic of an interaction, fostering a more competitive interaction between agents [64]. Even the mere risk of losing money to a robot could have elevated a participant's sense of uncertainty and perceived risk to a critical point, more so than experiments without monetary incentive. Future work with unobtrusive objective measures of threat perception (e.g., galvanic skin response, heart rate data) may serve to clarify the role of financial risk in user vulnerability.

Regarding our vulnerability categories 2, "When the robot asks for sensitive or personal information from the user and the user is unsure how that information is going to be used", and 4 "Interacting with a large, heavy robot in close proximity", our moderator analysis suggest that earning trust in these scenarios to be particularly challenging.

Studies included in vulnerability category 2 were unique in that the robot or agent involved tried to solicit either personal or private information from the participant. As discussed in the introduction, robots are particularly good at earning the trust of users [1], with other experiments showing that they can solicit personal information from the elderly [43], from people who have just been exposed to traumatic video footage [6], and also, that the type of information disclosed plays a part in participants trust evaluation [3]. Firstly, they often involved an initial phase of trust building followed by sensitive information requests [32, 37] or socially challenging information requests (e.g., to choose a alternate restaurant for a prospective customer) [19]. So, perhaps trust was harder to maintain because, in these type of experiments, the participants feel exposed because they are uncertain about how the information they have supplied is going to judged and handled. In experiments where trust was initially build, participants may have felt a sense of betrayal, as the robot disguised it's intentions to gather personal information as goodwill. Those who were happy to share their experiences with the robot must have judged that the ambivalent potential [34] of the interaction leaned more towards positive rather than a negative outcome. It would be interesting to see how these self-disclosure judgements vary in the population (the MDMT [35] scale could do this), as well as how the content of the conversation (e.g., something embarrassing or traumatic) interacts with this process. For now, our results seem to indicate that self-disclosure with a robot does not readily lead to increases in trust, indicating that the vulnerability experienced negatively impacts user's evaluations of the robot or agent.

Vulnerability category 4 relates specifically to the additional vulnerability point raised by Law and Scheutz [28], suggesting that large robots pose a unique physical danger to users. Whilst this summary might generalise people's impressions of large robots, a recent survey study has indicated that people consider large robots to be less safe [46]. It is perhaps not surprising therefore, that studies of trust including large robots failed to demonstrate a positive impact on user's trust. What is interesting about this finding, is that the studies included in this category were all collaborative. So, there is an argument to be made that trust should have been evidenced in this category given the robot's benevolence (e.g., [35]); to aid the user's efforts in completing a task. Supporting this point is the qualitative work by Hannibal [16], indicating that robotics experts feel that the risks posed by robots are understated by researchers, with the possible dangers involved frequently downplayed, e.g., that too few researchers acknowledged the physical risks robots posed. It may also explain why many authors in this domain employ Wizard of Oz type experiments, where the researcher controls the navigation and/or behaviour's

of the robot for safety reasons. In summary, it appears that the risk of physical harm posed by large, heavy robots makes the development of trust difficult. The vulnerability that individual's experience in these types of interactions have a negative impact on their assessment of the robot. Perhaps at this moment in time, we are not quite ready to be fully trusting of large robots, but our level of perceived threat (e.g., [65]) will adapt over time to accept them as suitable for certain roles; much like the way other large machinery came to be accepted during the industrial revolution [40].

A methodological reflection from our study is that vulnerability category 1 studies were the most popular in our sample. One way to explain this trend is that studies of performance risk are most often pursued in HRI research. This may be because of the challenges associated with the other risk categories. As Hannibal [16] demonstrated in their work, there is concern amongst robot researchers about studies that involve a degree of physical risk (in the present work this included the studies [27, 33, 50, 52]). Guidelines for HRI studies with the risk of harmful robot-to-human collisions suggest inclusion of a fail-safe mechanism to stop the robot, or to have an operator supervise and assume control of the robot if necessary (e.g., through a Wizard of Oz setup [61]). As such, the additional safety protocols and staff required for studies with a physical risk may encourage researchers to opt for safer HRI experiments. Moving forward, a set of guidelines for safe HRI research of trust would give researchers more confidence to pursue studies with a degree of physical risk, whilst also equipping university ethics boards with the knowledge to evaluate proposals fairly.

With respect to the implication of these findings in the context of HRT, we show that vulnerability, when operationalised, can offer some unique insights to trust development between humans and robots. Prior to our work, most studies in HRI provided a definition of trust built on the assumption that the trustor makes themselves vulnerable to the trustee for trust to be gained (e.g., see [29, 38]). Like Hannibal [16], we show that vulnerability is a multi-faceted concept that encapsulates perceived threat, issues with self-disclosure, and risk. Moving forward, we recommend that HRI researchers studying trust think carefully about their experimental scenario will affect their user's vulnerability. Is the robot threatening in any way? Does it suit the deployment setting? Do participants perceive the robot to be benevolent, or not? What can be done to make a large robot appear safer? These are just a few suggestions future designers and researchers should take into consideration.

On vulnerability and risk in HRI, we show how the two are related, and ultimately, the physical risk in HRI poses a future challenge for trust building. This is especially important, as hype often overshadows reality in popular discourse about robots [25] and AI (e.g., the threat posed by ChatGPT; [4]). The uneven number of studies and effect sizes across our vulnerability categories suggests more needs to be done, with emphasis on studies of *physical risk* in HRI. Further, many studies did not report adequate data for meta-analysis inclusion, and researchers should also be mindful of this going forward.

Presently, we are developing a survey in which naive robot users will evaluate the human trustor's vulnerability and situational risk from a series of recently conducted human-robot trust studies. In doing so, we will establish how people view the experiments being conducted in the field of HRI, in terms of the user's vulnerability and risks. The data will also implicitly indicate (by virtue of examining vulnerability and risk) scenarios that are threatening. Such information will help inform the design of future robots and AI, so that these technologies may be perceived more positively, where there were previous concerns from the public.

An important limitation to note is that our approach could be viewed as subjective rather than objective. The development of our vulnerability categories relied heavily on the descriptive reflections offered by Lewis et al. [30], Law and Scheutz [28], and Stuck et al. [55]. Because both vulnerability and risk are ill-defined concepts, we decided that the safest route to proceed in the development of our categorisation of vulnerability was to take inspiration from well-researched and informed review articles. And, whilst there were other definitions to choose from (e.g., Nienaber et al. [42] definition of vulnerability) we opted for the most contextually relevant works. In light of this, we recommend researchers in the field consider our categories as a starting point for further investigation of the vulnerabilities and risks people experience in HRI. Our aforementioned follow-up survey

study will glean further insights to people's perception of trustor vulnerabilities and the risks involved, validating our vulnerability categories empirically.

Furthermore, our process of vulnerability category assignment for our meta-analysis moderator variable could have been more methodologically and statistically rigorous. Recruiting researchers blind to the nature of the study to assign the categories to our study sample, and then calculating inter-rater reliability, would have improved the objectiveness of the vulnerability category assignment process. We will adopt these measures in future moderator meta-analysis work.

6 CONCLUSION

The present paper used a multilevel meta-analysis to determine the contribution of a new vulnerability variable to trust development from a set of identified studies. Our analysis indicated that multilevel moderator meta-analysis can be used to further probe the nature of trust in HRI, with studies of robot performance emerging as a unique phenomenon in trust development, relative to our other vulnerability categories. That is, the robot's ability to perform a task was the most likely to lead to gains in trust, relative to the other categories of vulnerability outlined.

The work marks an important theoretical development, whilst demonstrating the utility of multilevel moderator analysis for the examination of potentially relevant variables.

Broadly, the work suggests that vulnerability can be operationalised and deployed meaningfully as a predictor variable, despite its lack of coherence as a concept. In light of our findings, we recommend researchers in the field give careful consideration to user vulnerability in the design of their human-robot trust studies.

REFERENCES

[1] Takuto Akiyoshi, Junya Nakanishi, Hiroshi Ishiguro, Hidenobu Sumioka, and Masahiro Shiomi. 2021. A Robot That Encourages Self-Disclosure to Reduce Anger Mood. *IEEE Robotics and Automation Letters* 6, 4 (Oct. 2021), 7925–7932. https://doi.org/10.1109/LRA.2021.3102326

[2] D. D. Allan, Andrew J. Vonasch, and Christoph Bartneck. 2022. Better than Us: The Role of Implicit Self-Theories in Determining Perceived Threat Responses in HRI. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 215–224. https://doi.org/10.1109/HRI53351.2022.9889520

[3] Jessica K. Barfield. 2021. Self-Disclosure of Personal Information, Robot Appearance, and Robot Trustworthiness. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, Vancouver, BC, Canada, 67–72. https://doi.org/10.1109/RO-MAN50785.2021.9515477

[4] Luigi De Angelis, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferragina, Alberto Eugenio Tozzi, and Caterina Rizzo. 2023. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Frontiers in Public Health* 11 (2023), 1567.

[5] Cambridge Dictionary. 2023. *Cambridge Online Dictionary*. https://dictionary.cambridge.org/dictionary/english/vulnerability

[6] Yunfei (Euphie) Duan, Myung (Ji) Yoon, Zhixuan (Edison) Liang, and Johan Ferdinand Hoorn. 2021. Self-Disclosure to a Robot: Only for Those Who Suffer the Most. *Robotics* 10, 3 (July 2021), 98. https://doi.org/10.3390/robotics10030098

[7] Oxford Economics. 2019. *How Robotics Changed the World: What Automation Really Means for Jobs and Productivity*. Technical Report. Oxford Economics Ltd.

[8] Francesco Ferrari, Maria Paola Paladino, and Jolanda Jetten. 2016. Blurring Human–Machine Distinctions: Anthropomorphic Appearance in Social Robots as a Threat to Human Distinctiveness. *International Journal of Social Robotics* 8, 2 (April 2016), 287–302. https://doi.org/10.1007/s12369-016-0338-y

[9] Zuzana Fungáčová, Iftekhar Hasan, and Laurent Weill. 2019. Trust in banks. *Journal of Economic Behavior & Organization* 157 (2019), 452–476.

[10] Y Gao, E Sibirtseva, G Castellano, and D Kragic. 2019. Fast Adaptation with Meta-Reinforcement Learning for Trust Modelling in Human-Robot Interaction. *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 305–312. https://doi.org/10.1109/IROS40897.2019.8967924

[11] Ceenu George, Malin Eiband, Michael Hufnagel, and Heinrich Hussmann. 2018. Trusting Strangers in Immersive Virtual Reality. *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*. https://doi.org/10.1145/3180308.3180355

[12] Feyza Merve Hafızoğlu and Sandip Sen. 2018. Reputation Based Trust In Human-Agent Teamwork Without Explicit Coordination. *Proceedings of the 6th International Conference on Human-Agent Interaction, December 15-18*, 238–245. https://doi.org/10.1145/3284432.3284454

[13] Feyza Merve Hafızoğlu and Sandip Sen. 2019. Understanding the Influences of Past Experience on Trust in Human-Agent Teamwork. *ACM Trans. Internet Technol.* 19 (9 2019), 45–65. Issue 4. https://doi.org/10.1145/3324300

[14] PA Hancock, Theresa T Kessler, Alexandra D Kaplan, John C Brill, and James L Szalma. 2020. Evolving trust in robots: specification through sequential and comparative meta-analyses. *Human factors* (2020), 0018720820922080.

[15] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors* 53, 5 (2011), 517–527.

[16] Glenda Hannibal. 2021. Focusing on the Vulnerabilities of Robots through Expert Interviews for Trust in Human-Robot Interaction. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 288–293.

[17] Mathias Harrer, Pim Cuijpers, Furukawa Toshi A, and David D Ebert. 2021. *Doing Meta-Analysis With R: A Hands-On Guide* (1st ed.). Chapman & Hall/CRC Press.

[18] Mathias Harrer, Pim Cuijpers, Toshi Furukawa, and David Daniel Ebert. 2019. Dmetar: Companion R package for the guide'doing meta-analysis in R'. *R package version 0.0* 9000 (2019).

[19] Sarita Herse, Jonathan Vitale, Meg Tonkin, Daniel Ebrahimian, Suman Ojha, Benjamin Johnston, William Judge, and Mary Anne Williams. 2018. Do You Trust Me, Blindly? Factors Influencing Trust Towards a Robot Recommender System. *RO-MAN 2018 - 27th IEEE International Symposium on Robot and Human Interactive Communication* (2018), 7–14. https://doi.org/10.1109/ROMAN.2018.8525581

[20] Rens Hoegen, Aneja Deepali, Daniel McDuff, and Mary Czerwinski. 2019. An End-to-End Conversational Style Matching Agent. *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents : July 2-5, 2019*, 269. https://doi.org/10.1145/3308532.3329473

[21] Michiel Joosse and Vanessa Evers. 2017. A Guide Robot at the Airport: First Impressions. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Vienna Austria, 149–150. https://doi.org/10.1145/3029798.3038389

[22] Stephan Kambach, Helge Bruelheide, Katharina Gerstner, Jessica Gurevitch, Michael Beckmann, and Ralf Seppelt. 2020. Consequences of multiple imputation of missing standard deviations and sample sizes in meta-analysis. *Ecology and Evolution* 10 (10 2020), 11699–11712. Issue 20. https://doi.org/10.1002/ece3.6806

[23] Bing Cai Kok and Harold Soh. 2020. Trust in Robots: Challenges and Opportunities. *Current Robotics Reports* 1, 4 (Dec. 2020), 297–309. https://doi.org/10.1007/s43154-020-00029-y

[24] Johannes Maria Kraus, Florian Nothdurft, Philipp Hock, David Scholz, Wolfgang Minker, and Martin Baumann. 2016. Human after all: Effects of mere presence and social interaction of a humanoid robot as a co-driver in automated driving. *AutomotiveUI 2016 - 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Adjunct Proceedings*, 129–134. https://doi.org/10.1145/3004323.3004338

[25] Ingo Kregel, Julian Koch, and Ralf Plattfaut. 2021. Beyond the hype: robotic process automation's public perception over time. *Journal of organizational computing and electronic commerce* 31, 2 (2021), 130–150.

[26] Dean Langan, Julian P.T. Higgins, Dan Jackson, Jack Bowden, Areti Angeliki Veroniki, Evangelos Kontopantelis, Wolfgang Viechtbauer, and Mark Simmonds. 2019. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research Synthesis Methods* 10 (3 2019), 83–98. Issue 1. https://doi.org/10.1002/jrsm.1316

[27] Edith Law, Vicky Cai, Qi Feng Liu, Sajin Sasy, Joslin Goh, Alex Blidaru, and Dana Kulic. 2017. A Wizard-of-Oz Study of Curiosity in Human-Robot Interaction. *In 2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), August 28-Sept 1*, 607–614.

[28] Theresa Law and Matthias Scheutz. 2021. *Chapter 2 - Trust: Recent concepts and evaluations in human-robot interaction.* Academic Press, 27–57. https://doi.org/10.1016/B978-0-12-819472-0.00002-2

[29] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[30] Michael Lewis, Katia Sycara, and Phillip Walker. 2018. The Role of Trust in Human-Robot Interaction. In *Foundations of Trusted Autonomy*, Hussein A. Abbass, Darryn J. Reid, and Jason Scholz (Eds.). Springer, Cham, Switzerland, Chapter 8, 135–159.

[31] Yuhua Liang and Seungcheol Austin Lee. 2016. Advancing the Strategic Messages Affecting Robot Trust Effect: The Dynamic of User-and Robot-Generated Content on Human-Robot Trust and Interaction Outcomes. *Cyberpsychology, Behavior, and Social Networking* 19 (2016), 538–544. Issue 9. https://doi.org/10.1089/cyber.2016.0199

[32] Christine Lisetti, Reza Amini, Ugan Yasavur, and Naphtali Rishe. 2013. I can help you change! An empathic virtual agent delivers behavior change health interventions. *ACM Transactions on Management Information Systems* 4 (2013), 1–28. Issue 4. https://doi.org/10.1145/2544103

[33] Shih Yun Lo, Elaine Schaertl Short, and Andrea L. Thomaz. 2020. Planning with partner uncertainty modeling for efficient information revealing in teamwork. *ACM/IEEE International Conference on Human-Robot Interaction* (2020), 319–327. https://doi.org/10.1145/3319502.3374827

[34] Catriona Mackenzie. 2020. Vulnerability, Insecurity and the Pathologies of Trust and Distrust. *International Journal of Philosophical Studies* 28, 5 (2020), 624–643. https://doi.org/10.1080/09672559.2020.1846985 arXiv:https://doi.org/10.1080/09672559.2020.1846985
[35] Bertram Malle and Daniel Ullman. 2020. A multidimensional conception and measure of human-robot trust. In *Trust in Human-Robot Interaction*, Chang Nam and Joseph Lyons (Eds.). Academic Press, Chapter 1, 3–25.
[36] Jordan A. Mann, Bruce A. Macdonald, I. Han Kuo, Xingyan Li, and Elizabeth Broadbent. 2015. People respond better to robots than computer tablets delivering healthcare instructions. *Computers in Human Behavior* 43 (2015), 112–117. https://doi.org/10.1016/j.chb.2014.10.029
[37] Nikolas Martelaro, Victoria C. Nneji, Wendy Ju, and Pamela Hinds. 2016. Tell me more: Designing HRI to encourage more trust, disclosure, and companionship. *ACM/IEEE International Conference on Human-Robot Interaction* 2016-April, 181–188. https://doi.org/10.1109/HRI.2016.7451750
[38] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.
[39] Frauke Meyer, Deidre M Le Fevre, and Viviane MJ Robinson. 2017. How leaders communicate their vulnerability: Implications for trust building. *International Journal of Educational Management* (2017).
[40] Haradhan Kumar Mohajan. 2019. The First Industrial Revolution: Creation of a New Global Human Era. *Journal of Social Sciences and Humanities* 5, 4 (2019), 377–387.
[41] Manisha Natarajan and Matthew Gombolay. 2020. Effects of Anthropomorphism and Accountability on Trust in Human Robot Interaction. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Cambridge United Kingdom, 33–42. https://doi.org/10.1145/3319502.3374839
[42] Ann-Marie Nienaber, Marcel Hofeditz, and Philipp Daniel Romeike. 2015. Vulnerability and trust in leader-follower relationships. *Personnel Review* (2015).
[43] Yohei Noguchi, Hiroko Kamide, and Fumihide Tanaka. 2020. Personality Traits for a Social Mediator Robot Encouraging Elderly Self-Disclosure on Loss Experiences. *ACM Transactions on Human-Robot Interaction* 9, 3 (Sept. 2020), 1–24. https://doi.org/10.1145/3377342
[44] Mourad Ouzzani, Hossam Hammady, Zbys Fedorowicz, and Ahmed Elmagarmid. 2016. Rayyan—a web and mobile app for systematic reviews. *Systematic reviews* 5, 1 (2016), 1–10.
[45] R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
[46] Matteo Rubagotti, Inara Tusseyeva, Sara Baltabayeva, Danna Summers, and Anara Sandygulova. 2022. Perceived safety in physical human–robot interaction—A survey. *Robotics and Autonomous Systems* 151 (May 2022), 104047. https://doi.org/10.1016/j.robot.2022.104047
[47] Nicole Salomons, Michael van der Linden, Sarah Strohkorb Sebo, and Brian Scassellati. 2018. Humans Conform to Robots: Disambiguating Trust, Truth, and Conformity. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 187–195. https://doi.org/10.1145/3171221.3171282
[48] Tracy. L. Sanders, Tarita. Wixon, K E. Schafer, Jessie Y.C. Chen, and Peter A. Hancock. 2014. The Influence of Modality and Transparency on Trust in Human-Robot Interaction. *2014 IEEE International Inter-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, 156–159. https://doi.org/10.1109/CogSIMA.2014.6816556
[49] Sarah Strohkorb Sebo, Margaret Traeger, and Brian Scassellati. 2018. The Ripple Effects of Vulnerability : The Effects of a Robot ' s Vulnerable Behavior on Trust in Human-Robot Teams. *In Proceedings of2018 ACM/IEEE International Conference on Human-Robot Interaction, March 5-8*, 178–186. https://doi.org/10.1145/3171221.3171275
[50] Booth Serena, Tompkin James, Pfister Hanspeter, Waldo Jim, Gajos Krzysztof, and Nagpa Radhika. 2017. PiggybackingRobots: Human-Robot Overtrust in University Dormitory Security. *ACM/IEEE International Conference on Human-Robot Interaction, March 06-09*, 426–434.
[51] Sharon Shavitt and Aaron J. Barnes. 2020. Culture and the Consumer Journey. *Journal of Retailing* 96, 1 (2020), 40–54. https://doi.org/10.1016/j.jretai.2019.11.009 Understanding Retail Experiences and Customer Journey Management.
[52] Vasant Srinivasan, Leila Takayama, and Willow Garage. 2016. Help Me Please : Robot Politeness Strategies for Soliciting Help From People. *CHI '16, May 07–12*, 4945–4955. https://doi.org/10.1145/2858036.2858217
[53] Christopher John Stanton and Catherine J Stevens. 2017. Don't stare at me: the impact of a humanoid robot's gaze upon trust during a cooperative human–robot visual task. *International Journal of Social Robotics* 9 (2017), 745–753.
[54] Julia G. Stapels and Friederike Eyssel. 2022. Robocalypse? Yes, Please! The Role of Robot Autonomy in the Development of Ambivalent Attitudes Towards Robots. *International Journal of Social Robotics* 14, 3 (April 2022), 683–697. https://doi.org/10.1007/s12369-021-00817-2
[55] Rachel E. Stuck, Brittany E. Holthausen, and Bruce N. Walker. 2021. Chapter 8 - The role of risk in human-robot trust. In *Trust in Human-Robot Interaction*, Chang S. Nam and Joseph B. Lyons (Eds.). Academic Press, 179–194. https://doi.org/10.1016/B978-0-12-819472-0.00008-3
[56] Rüdiger M Trimpop. 1994. *The psychology of risk taking behavior*. Elsevier.
[57] Cas D. P. Van'T Hullenaar, Paula Bos, and Ivo A. M. J. Broeders. 2019. Ergonomic assessment of the first assistant during robot-assisted surgery. *Journal of Robotic Surgery* 13, 2 (April 2019), 283–288. https://doi.org/10.1007/s11701-018-0851-0

[58] Wolfgang Viechtbauer. 2010. Conducting Meta-Analyses in *R* with the 'metafor' Package. *Journal of Statistical Software* 36 (2010). Issue 3. https://doi.org/10.18637/jss.v036.i03

[59] Wolfgang Viechtbauer and Mike W-L Cheung. 2010. Outlier and influence diagnostics for meta-analysis. *Research synthesis methods* 1, 2 (2010), 112–125.

[60] Alan R Wagner, Paul Robinette, and Ayanna Howard. 2018. Modeling the human-robot trust phenomenon: A conceptual framework based on risk. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 4 (2018), 1–24.

[61] Michael L Walters, Sarah Woods, Kheng Lee Koay, and Kerstin Dautenhahn. 2005. Practical and methodological challenges in designing and conducting human-robot interaction studies. In *Proceedings of the AISB 05 Symposium on Robot Companions*. AISB.

[62] A Wong, A Xu, and G Dudek. 2019. Investigating Trust Factors in Human-Robot Shared Control: Implicit Gender Bias Around Robot Voice. *2019 16th Conference on Computer and Robot Vision (CRV)*, 195–200. https://doi.org/10.1109/CRV.2019.00034

[63] Kumar Yogeeswaran, Jakub Złotowski, Megan Livingstone, Christoph Bartneck, Hidenobu Sumioka, and Hiroshi Ishiguro. 2016. The Interactive Effects of Robot Anthropomorphism and Robot Ability on Perceived Threat and Support for Robotics Research. *Journal of Human-Robot Interaction* 5, 2 (Sep 2016), 29. https://doi.org/10.5898/JHRI.5.2.Yogeeswaran

[64] Debora Zanatto, Massimiliano Patacchiola, Jeremy Goslin, Serge Thill, and Angelo Cangelosi. 2020. Do Humans Imitate Robots? An Investigation of Strategic Social Learning in Human-Robot Interaction. *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 449–457. https://doi.org/10.1145/3319502.3374776

[65] Jakub Złotowski, Kumar Yogeeswaran, and Christoph Bartneck. 2017. Can we control it? Autonomous robots threaten human identity, uniqueness, safety, and resources. *International Journal of Human-Computer Studies* 100 (April 2017), 48–54. https://doi.org/10.1016/j.ijhcs.2016.12.008

7 APPENDIX

7.1 R code for Meta-analysis model

All R scripts can be found on Open Science Framework (OSF).

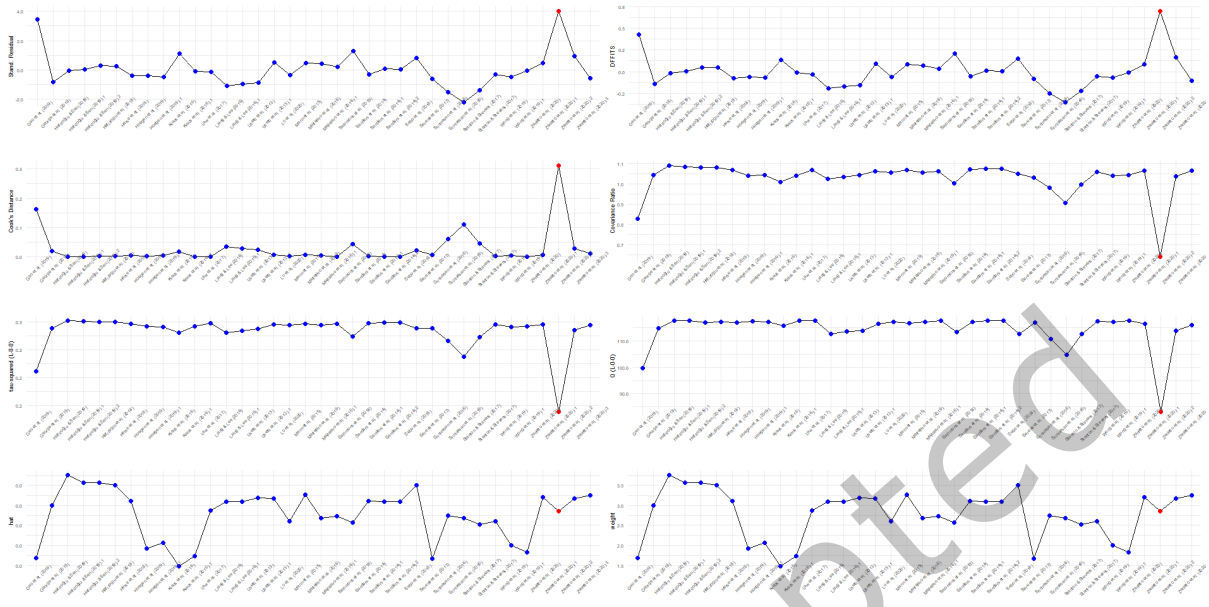

Fig. 4. Influence diagnostics.

Fig. 5. Baujat Plot.

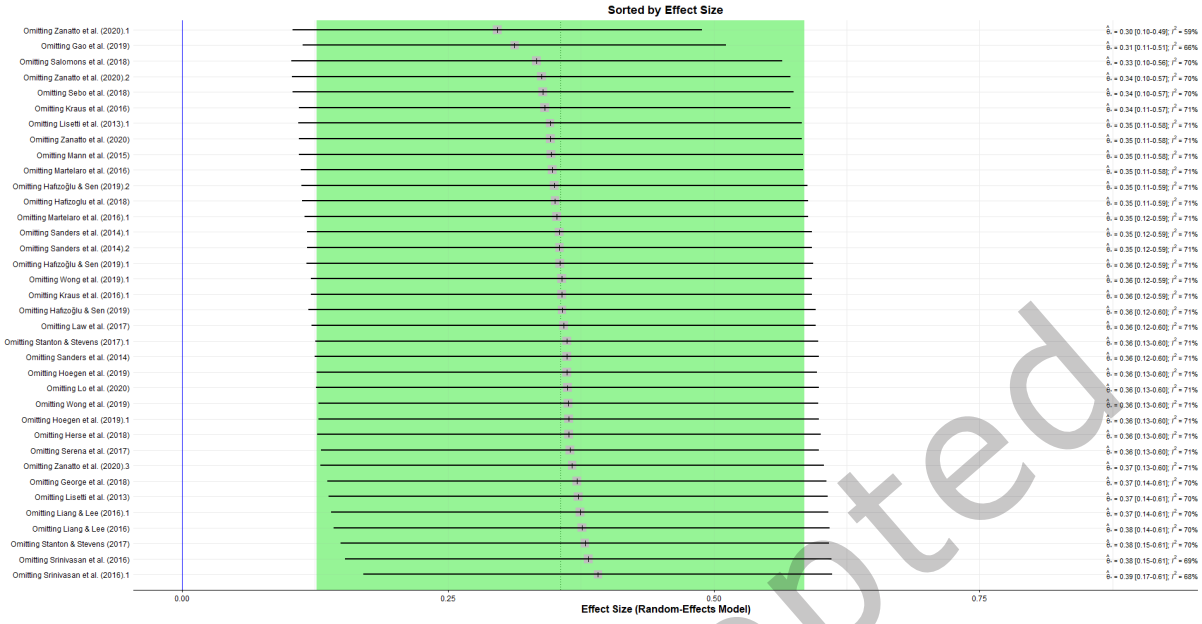

Fig. 6. Effect size omissions.

Fig. 7. I^2 variance from effect size omission.