15 Years of (Who)man Robot Interaction: Reviewing the H in Human-Robot Interaction

Katie Winkle, Erik Lagerstedt, Ilaria Torre, Anna Offenwanger

15 Years of (Who)man Robot Interaction: Reviewing the H in Human-Robot Interaction

KATIE WINKLE*, Uppsala Universitet, Sweden
ERIK LAGERSTEDT*, University of Skövde, Sweden
ILARIA TORRE*, KTH Royal Institute of Technology, Sweden
ANNA OFFENWANGER*, Université Paris-Saclay, CNRS, Inria, LISN, France

Recent work identified a concerning trend of disproportional gender representation in research participants in Human-Computer Interaction (HCI). Motivated by the fact that Human-Robot Interaction (HRI) shares many participant practices with HCI, we explored whether this trend is mirrored in our field. By producing a dataset covering participant gender representation in all 684 full papers published at the HRI conference from 2006-2021, we identify current trends in HRI research participation. We find an over-representation of men in research participants to date, as well as inconsistent and/or incomplete gender reporting which typically engages in a binary treatment of gender at odds with published best practice guidelines. We further examine if and how participant gender has been considered in user studies to date, in-line with current discourse surrounding the importance and/or potential risks of gender based analyses. Finally, we complement this with a survey of HRI researchers to examine correlations between the who is doing with the who is taking part, to further reflect on factors which seemingly influence gender bias in research participation across different sub-fields of HRI. Through our analysis we identify areas for improvement, but also reason for optimism, and derive some practical suggestions for HRI researchers going forward.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**.

Additional Key Words and Phrases: Gender, Systematic Review, User Study Methodologies, Participant Recruitment, Inclusivity

ACM Reference Format:
Katie Winkle*, Erik Lagerstedt*, Ilaria Torre*, and Anna Offenwanger*. 2022. 15 Years of (Who)man Robot Interaction: Reviewing the H in Human-Robot Interaction. *ACM Trans. Hum.-Robot Interact.* , (November 2022), 28 pages. https://doi.org/10.1145/3571718

1 INTRODUCTION

Since the start of the IEEE/ACM International Conference on Human-Robot Interaction in 2006, (some) researchers have been concerned with whether user gender might influence human-robot interaction (HRI) [26, 32, 88]. Today, HRI works continue to examine the impact of user gender [52, 69, 78], robot gendering [15, 93] and if/how these two might interact [25, 34, 45, 58]. At the same time, recent critiques have drawn attention to the way that current approaches to technology development and deployment may be upholding and reinforcing historical systems of gender-based oppression. This can happen through e.g., subtly favouring specific gender identities in recruitment and software development [17, 55], but also through data exclusion [62, 70, 82, 87], embedding

*All authors contributed equally to this work. KW has taken on first author responsibilities whilst 2nd-4th author ordering was decided by dice roll.

Authors' addresses: Katie Winkle*, Uppsala Universitet, Sweden, katie.winkle@it.uu.se; Erik Lagerstedt*, University of Skövde, Sweden, erik.lagerstedt@his.se; Ilaria Torre*, KTH Royal Institute of Technology, Sweden, ilariat@kth.se; Anna Offenwanger*, Université Paris-Saclay, CNRS, Inria, LISN, Orsay, France, anna.offenwanger@universite-paris-saclay.fr.

gender bias directly into machines [13, 70, 77] and designing technologies which propagate harmful gender norms [37, 84, 90].

A number of ethical AI and robotics guidelines explicitly identify the need for diversity regarding who gets to contribute to the design and evaluation of autonomous systems. For example, the Foundation for Responsible Robotics identifies that responsible robotics starts during research and development, and includes *'ensuring that a diverse set of viewpoints are represented in the development of the technology'* [1]. Similarly, the IEEE Ethically Aligned Design identifies that the risk of developing systems which disadvantage specific groups can be addressed by *'including members of diverse social groups in both the planning and evaluation of AI systems'* [2]. In addition to this ethical imperative, and specifically relating to gender, Tannenbaum et al. argue that sex and gender analyses can *"foster scientific discovery, improve experimental efficiency and enable social equality"* and have called on researchers to coordinate their efforts to implement robust methods of sex and gender analysis [85]. This call was echoed again by one of Tannenbaum's co-authors, Friederike Eyssel, in her keynote at the 2022 IEEE/ACM Conference on Human-Robot Interaction [29]. To summarise: who is taking part in HRI research concerns us not only because we want robust and generalisable results, but also because diversity in our user studies is a crucial precursor to the ethical design and development of effective HRI.

It is important to note that diversity in research participation is a complex and multifaceted concept. Dimensions of diversity include (but are not limited to) age, race, nationality, language, ableness, socio-economic status, and of course, gender. Intersectionality is the term used to describe how these dimensions of diversity can interact, and has been examined in HCI research [73]. This previous research shows that while we see some reporting for multiple dimensions of diversity, gender is by far the most common to be reported along with age. This is likely due to the fact that it is standard practice in human subject research to report participant gender (see e.g. APA reporting guidelines[3]) for reasons pertaining to e.g., generalisability and the support of meta-analyses. As such, while it would be preferable to examine participant diversity more holistically and intersectionally, in this work we collect data on gender alone. However, we do consider some implications of intersectionality (specifically gender x academic discipline) when examining who is *doing* HRI. In the field of computing, a reduced sense of belonging (correlated with increased dropout rate) has been identified among people with sexual orientations or gender identities that do not conform with stereotypes in the field [83], so we might expect a similar pattern of identity within HRI.

In their 2021 paper at the ACM Conference on Human Factors in Computing Systems (CHI), Offenwanger et al. identified a number of concerning trends regarding gender bias (see definitions under Section 1.2) in Human-Computer Interaction (HCI) research [59]. Undertaking a systematic review of 1,147 CHI papers published between 1981 and 2020, the authors documented a persistent under-representation of women (including a steady decline in the number of women participating in studies hosted on Amazon Mechanical Turk), as well as the invisibility and othering of non-binary participants. They also noted that gender bias patterns vary across sub-topics of HCI research with e.g. studies pertaining to physical interaction and virtual environments having lower representation of women than those pertaining to family and home or community infrastructure. Given that HRI is known to employ similar methods to HCI when it comes to design and user studies [40, 51, 63], Offenwanger et al.'s findings motivate a thorough reflection on *how* we are doing HRI, perhaps more specifically *who* we are inviting to do it with us, and how we are reporting that in our publications.

[1] https://responsiblerobotics.org/mission/

[2] https://standards.ieee.org/industry-connections/ec/ead-v1.html

[3] see APA7 Section 5.5: https://apastyle.apa.org/jars

With 2021 marking 15 years of the annual IEEE/ACM Conference on Human-Robot Interaction, we take this opportunity to investigate research trends and practices documented in works accepted for publication at the conference to date, as a snapshot of the HRI field (and its development over time) more broadly. We replicate and expand on Offenwanger et al.'s process to review the 684 papers published at the HRI conference from its inception in 2006 up until 2021, and report on the proportions of men, women and non-binary participants included in research published there to date. We additionally reflect on different practices to quantify diversity of representation, inspired by domains such as ecology and applied as measures for diversity among participants of conferences and experiments, noting the tension of attempting to capture something fundamentally complex in one comparable number. Motivated by Tannenbaum et al. [85], we further comment on if/how researchers in our community have been treating and/or examining the role of gender in HRI with respect to their data analyses. We make the dataset resulting from our systematic review available for the community [4]. Finally, we complement this systematic data collection and review of the conference literature with a survey of HRI researchers to examine any connections between *who is doing* and *who is taking part* in HRI research. This is done by dividing HRI research into sub-topics, and looking at trends within those in terms of participants' gender and researchers' gender, academic background and current research field.

1.1 Article Overview

This article paper is structured as follows. We first aim to establish a common ground with the reader by defining what is meant with the words "gender", "sex", and "bias" in the current paper (Section 1.2) before stating our research questions (Section 1.3). In Section 2 we then present related work on the topics of gender, diversity and critique of HRI research practice. Section 3 describes the research participation dataset we created (and how we did so) in order to examine research participation in studies published at the HRI conference to date. Section 4 presents the complementary survey we administered to HRI researchers in order look for any links between researcher identity, HRI sub-topic and research participation. Our findings are presented, with signposting to our research questions, in Section 5. We discuss key implications from these findings in Section 6 resulting in a summary of practical suggestions (at the individual researcher and research community levels) in Section 6.3. Finally, we provide a brief summary conclusion in Section 7.

1.2 Definitions of Gender, Sex and Bias

Bias is a word that has many meanings in different fields, so for clarity, we here outline what we mean by gender, and gender bias. When defining gender, it is important to note that sex and gender are separate things, but frequently conflated. Offenwanger et al. [59], referred readers to the definition provided by the Canadian Institutes of Health Research, which clearly differentiates the two: *"Sex refers to a set of biological attributes in humans and animals. It is primarily associated with physical and physiological features including chromosomes, gene expression, hormone levels and function, and reproductive/sexual anatomy. [...] Gender refers to the socially constructed roles, behaviours, expressions and identities of girls, women, boys, men, and gender diverse people. It influences how people perceive themselves and each other, how they act and interact, and the distribution of power and resources in society"* [1]. This is consistent with definitions from research institutes in other governments such as Sweden [3], where the majority of the authors of this paper are currently based. In this work, we refer to gender and gender representation, rather than sex, but acknowledge that (a) sex is also relevant, as sex-related differences (that in some cases might correlate with gender) can impact research results and (b) researchers might be conflating sex

[4]https://osf.io/xw7m9/

and gender when reporting participant demographics, exasperated by the fact that so few articles specify how questions surrounding gender and/or sex were asked (see related commentary under Section 6.2).

As per Offenwanger et al. [59], we knew that it was unlikely we would be able to identify exactly how different authors were operationilising gender, i.e., did they ask participants about gender, or sex? How was this question asked, if multiple choice, what answer options were given? Previous work on gender sensitivity (or a lack thereof) in HCI research has previously called attention to this issue, and called for researchers to be upfront in their understanding of gender when writing about it [18]. In our case, we also want to state our definition of gender before we are faced, in the literature we review, with conceptions of gender that might be incompatible with it, because if we don't, our paper reinforces a conception of gender that we do not agree with. As such, we also extend our definition of gender to include that *"[g]ender identity is not confined to a binary (girl/woman, boy/man) nor is it static; it exists along a continuum and can change over time. There is considerable diversity in how individuals and groups understand, experience and express gender through the roles they take on"* again per the Canadian Institutes of Health Research [1] and again in line with information from the national health service in Sweden [5]. Whilst some fields typically group gender variables with sexual orientation under the umbrella term Sexual Orientation and Gender Identity (SOGI), we saw no evidence of this in the HRI literature and limit our discussions here to sex and gender only.

We want to explicitly affirm our viewpoint that gender goes beyond binary, even though we have not always been able to treat it as such in this work due to the historically binary assumption/treatment of gender evidenced in those papers we've reviewed, and in (some of) the diversity measures we have replicated (i.e. Offenwanger et al.'s Distance from Even Representation measure [59], see Section 2). The binary assumption of gender exists both in government systems [37, 75] and in research, and causes problems for many reasons, among them the difficulty of tracking the participation of non-binary people in research [59]. While this binary treatment is now widely acknowledged to be problematic, and steps are being taken to rectify it legally [2, 3, 43, 76, 79, 80] and in research [50, 71], this is a slow process, and we are still confronted with binary treatments of gender.

With regard to the definition of gender bias, previous definitions include: *"any set of attitudes and/or behaviors which favors one sex over the other"* [10, p. 83] and: *"a systematically erroneous gender dependent approach related to social construct, which incorrectly regards women and men as similar/different"* [66, p. ii46]. In the first instance then, when we speak of gender bias in research participation we are talking about systems and practices that result in an over-representation of one gender over others. The latter definition alludes then to biases arising from failing to account for gender differences, *or* from looking for (and perhaps post-hoc rationalising) gender differences where one should not — a theme we discuss in the context of when (not) to do gender analyses and the seemingly divergent views HRI researchers hold on this topic.

1.3 Research Questions

This article aims to investigate gender representation and diversity in HRI research participation and factors which might influence it. To this end we address the following questions, via a systematic review of HRI literature and a survey of HRI researchers. The literature review represents an (extended) conceptual replication of Offenwanger et al.'s HCI work [59]. Based on their findings and resultant hypotheses, we examine variations in research participation across sub-topics of HRI and reported participant recruitment practices, both of which Offenwanger et al. linked with

[5]https://www.1177.se/liv--halsa/konsidentitet-och-sexuell-laggning/konsidentitet-och-konsuttryck/

variation in gender bias. The survey adds value to this review, first and foremost, by allowing for examination of Offenwanger et al.'s hypothesis that variation in researcher identity might correlate with diversity in research participation.

RQ1 Is there evidence of gender bias (per Section 1.2) in HRI research participation to date?
RQ2 (How) does any such bias vary across:
  a. different sub-topics (identified using probablistic topic modelling per [59]) within HRI ?
  b. different participant recruitment practices?
RQ3 Who, in terms of gender and educational background, is working on the different sub-topics of HRI that we identify within the literature?

and, based on the above:

RQ4 Is there any link between researcher identity and participant gender diversity/representation?

We further take the opportunity to reflect on what HRI researchers are doing with gender data, and what (other) data by asking:

RQ5 How many papers document analyses and/or report effects relating to participant gender? Of these, how many such papers were primarily concerned with participant gender as an independent experimental variable of interest, versus a potential confound?

Together, these questions represent a thorough reflection on the *who, how and why* of HRI user studies, which are a cornerstone of HRI research. As we discuss in more detail later (Section 5), the majority of HRI conference papers each year report some sort of user study, by which we mean any study including human subjects, therefore including e.g. design studies and basic usability testing.

2 RELATED WORK

2.1 Gender and Diversity in Human-Robot Interaction

Relatively few works in HRI explicitly set out to investigate how user gender may influence HRI: only 31 of the 684 papers we reviewed were identified as addressing explicit research question(s) and/or testing specific hypotheses regarding user gender. However, a larger number of papers (n = 64) were identified as including gender in their analyses e.g. as a confounding variable. Across all of these works, just over half (n = 50) reported a significant effect of user gender with the remaining (n = 45) reported the lack of such (see full results under Section 5.3). Many such analyses pertain to whether perceptions of a particular robot, robot behaviour or application vary across participant gender (e.g. [5, 21, 68, 72]) but others pertain to behavioural differences (or lack thereof) when interacting with a robot (e.g. [30, 31, 54, 74]). As such, previous literature paints a mixed picture with regards to if, when, and how a person's gender might *directly* impact their experience of HRI in the context of these HRI experimental settings and measures.

Given the relatively small overall number of papers discussing gender in our corpus (a total of 95 out of 684) as well as potentially divergent views within the HRI community (see further comments on this under Section 6.2.1) a broader question might be: (when) is gender analysis appropriate in the first place? As previously noted, Tannenbaum et al. claim that, per the title of their Nature perspectives article, 'sex and gender analysis improves science and engineering' [85]. The authors draw attention to the necessity of gender and sex analysis for equality e.g. in designing for physical safety (car safety devices which account for sex differences in physical size) and reducing gender bias in AI systems (persons in a kitchen being default tagged as women by image labelling systems). Notably however, on this last point, they note that *"a first challenge in algorithmic bias is to identify when it is appropriate for an algorithm to use gender information. In some settings, such as the assignment of job ads, it might be desirable for the algorithm to explicitly ignore the gender of an individual as well as features such as weight, which may correlate with gender but are not*

directly related to job performance.". However, others might argue that gender *should* explicitly be considered in the context of job applications (although arguably by trained *human* recruiters, rather than any sort of algorithmic system) as a way to account for effects of gender bias in the workplace [65] and/or to support affirmative action policies which work towards diversity goals of an organisation [86, ch. 7].

This importance of *context* over *categorisation* is taken up in another recent Nature perspectives article which calls for researchers to move away from social categories (including e.g. race and gender) and instead consider social context [23]. The authors Cikara et al. note that *"gender and racial or ethnic categories are often treated as purposive social groups, as though tagging a person with a category comprehensively capture their thoughts, feelings and perceptions [...] Repeatedly relying on categories in research gives rise to illusory essences — the notion (even among experts) that these categories represent objective, definable and fixed constructs — which in turn, reifies the categories.".* A better question then might be: how can we conduct gender sensitive research in HRI? Here we can take inspiration from our colleagues in HCI who have a head start on us in critically considering these issues, and we want to spotlight Burtscher and Spiel's article on this topic as an excellent starting point [18].

Regardless of the above, diversity in research participation is important from both a technical and ethical perspective with regards to delivering effective HRI that has positive impact in the real-world. As summarised by Offenwanger et al. [59], failure to ensure diversity (with respect to gender, in the first instance) can have a negative impact on underrepresented populations *"through unintentional exclusion [62, 87], which can lead to role stereotyping [13, 77] and new technologies being difficult to use for underrepresented groups [4, 16, 53, 56, 70, 82]".* Under-representation can be especially problematic for people with non-normative gender identities, as exclusion can be easily caused by inadequate analysis tools. Most gender analyses, particularly statistical analyses, are confined to only analysing men and women, and it is not unusual for (even the most well-intentioned) researchers to exclude participants with non-normative identities due to sample sizes being too low for inclusion in statistical tests. Even the *Distance from Even Representation* (DER) measure used by Offenwanger et al. to illustrate the under-representation of women in HCI studies *"is confined to comparing the representation of men and women and cannot be used to analyse more complex gender representation"* [59]. Alternative metrics such as diversity measures from ecology [33, 41] have been proposed as a way to address this problem, and we discuss the practical applicability of these under Section 2.3.

2.2 Previous Critique of HRI (and Related) Research Practice

HRI researchers have noted the importance of conducting user and design studies with diverse sets of participants that reflect the eventual target population [14, 67]. The publications in the first seven editions of the HRI conference had, however, an over-representation of men among their participants [89]. To our knowledge, this is the only equivalent previous work specifically concerned with examining gender representation in HRI research participants. However, HRI researchers have been very active in critiquing the field's typical research practices. For example, work has been done on examining author diversity [20], classifying HRI publication methodology [9], reviewing data collection tools [22, 74], calling out the need for replication of studies [44], and pointing out a troubling lack of reporting with regards to ethics practices [64], recruitment, compensation, and participant gender [24].

Moving from HRI to our neighbouring field of HCI, previous works have examined researcher diversity [11, 41, 92] and its relation to publishing [57], evaluation methods in HCI [7], participant sample sizes [19], participant gender [18, 59, 71], and intersectionality [18, 73]. Existing guidelines for collecting and reporting participant gender [71] and commentary on how researchers might

practice (gender) sensitive research [18] are essential reading for those in HRI interested in this topic, and very much inform the recommendations we later present under Section 6.3. In the fields of computer science and AI more broadly, initiatives like the UNESCO report on the digital gender divide [90] and UNICEF's policy guidance on the development of AI for and with children [28] further call for gender diversity in those involved with the development and evaluation of AI-enabled systems.

2.3 Quantifying Diversity in Research

Diversity is notoriously difficult to define, partly due to its close relation to complexity, and to the hierarchical interdependence and fuzziness of many categories [49, 60]. Although no definition is perfect there are many useful ways to define diversity in measurable ways, whether within or between categories.

To facilitate monitoring of diversity and representation in scientific communities several ways of quantifying diversity have been considered. Inspired by ecology, a *Gender Diversity Index* (GDI) has been proposed [33, 41], in which weighted Shannon indices [81] were calculated to measure gender diversity of keynote speakers, authors and conference organisers. The proposed benefit of the GDI is that one number in the range [0;1] will capture several aspects of diversity, where a higher number corresponds to better diversity, making comparison and tracking over time easier. The boundaries are achieved by normalising the Shannon index H', that is, calculating a Pielou index J';

$$J' = \frac{H'}{H'_{max}} = -\frac{1}{\ln S} \sum_{i=1}^{S} p_i \ln p_i, \tag{1}$$

in which p_i is the relative abundance of species i, and S is the number of species. To capture more aspects of diversity Hupont et al. [41] created a Conference Diversity Index consisting of the average of the three Pielou indices for gender, institute affiliation and geographic provenance.

In ecology there is a plethora of diversity measures, and many of them can be written in the form of a Hill number,

$$^{q}D \equiv \left(\sum_{i=1}^{S} p_i^q \right)^{1/(1-q)} \tag{2}$$

in which p_i is the relative abundance of species i, and q is a constant defining different diversity measures [47]. At the limit when q approaches 1, the diversity measure becomes the exponential of the Shannon index. The diversity can be measured within a community (alpha diversity), between communities (beta diversity), or a single measure can be calculated to capture both (gamma diversity) [48]. However, the use of diversity measures in ecology is contested [39]. The simplicity of such indices makes comparisons easy, but the lack of context or depth can make the comparisons unfair or misleading. After all, the clarity of the index is a consequence of a lot of information being excluded by the compression. The various indices promote different aspects of diversity, such as number of species (richness), the relative abundance of the species (evenness), or presence of rare or endangered species. The Shannon index depends on both the richness and the evenness, but the Pielou index is achieved by discarding the richness.

The diversity measures mentioned thus far all rely on entropy to conceptualise diversity. An alternative, for instance used by Offenwanger et al. [59] in their *Distance from Even Representation* (DER) measure, is to look at distance. The DER measure was specifically developed to measure inequality of representation between men and women, defined as

$$DER = \frac{women - men}{women + men}. \tag{3}$$

This results in a number ranging from [-1;1] where 0 corresponds to equal representation, negative values to an over-representation of men, and positive values to an over-representation of women. Although this metric lacks the ability to handle more than two genders, or intersectional aspects of gender, its simplicity makes both interpretations and limitations clear.

In the current paper, we report both DER and GDI measures from our collected data, to allow a direct comparison to Offelwanger et al. [59]'s work on the CHI and Hupont et al. [41]'s work on the ACII (Affective Computing and Intelligent Interaction) communities. However, we are aware that none of these measures are well suited to capture the complexities of gender spectra and dynamics. We would therefore like to take this opportunity to call for novel and more appropriate measures of diversity. We would at the same time point out that all aspects of diversity cannot be captured by an index, so when using a diversity measure it is important to complement it with a discussion about what aspects that specific measure highlights, and how to appropriately interpret the results.

3 THE HRI RESEARCH PARTICIPATION DATASET

We annotated the 684 full papers published at the ACM/IEEE HRI conference[6] from 2006 to 2021 (excluding extended abstracts, LBR's, student design competitions, and video submissions). Works at this conference have previously been studied to identify trends and traditions in HRI (e.g. [9, 74, 89]), and it is arguably the closest HRI equivalent to the ACM SIGCHI conference in HCI, as used by Offenwanger et al. [59]. Further, the highly selective nature of the conference (the acceptance rate has consistently remained at, or just below, 25% for more than 10 years) implies that work accepted for publication is considered high quality by the community. It seems sensible to assume, on this basis, that methodologies and practices documented in the conference works are likely to be seen as good practice and propagate through the field more broadly. The conference's specific focus on HRI, while accepting a wide variety of approaches to the field, makes ACM/IEEE HRI somewhat special compared to other robotics conferences that tend to have more of a technical focus and often relegate HRI to specific sub-tracks. For these reasons, we expect a wide variety also in terms of awareness and approaches with respect to participant gender, expressed both as presence of good examples as well as a lack of consensus or awareness regarding best practice.

3.1 Data Collection Tool and Data Schema

We contacted the authors of Offenwanger et al. [59], who provided us with a copy of the Machine Assisted Gender Data Annotation (MAGDA) tool for this analysis. The MAGDA tool was designed to complement the data schema developed by Offenwanger et al., which we adhered to in our data collection. In this way, our dataset (which we release with this publication) complements and extends the dataset released with their publication. In addition to the data captured by Offenwanger et al. (number of participants and participant information, including both demographics and recruitment practices), we also extracted text that relates to analysis of participant gender (this would typically be found in the discussion or results section, e.g. *"we used gender of the participants as a control variable"*). 10% of the data (70 papers, randomly selected) were annotated by four coders to calculate inter-rater agreement. Agreement was high for overall number of participants (ICC = 0.97), and number of female and male participants (ICC = 0.84 and 0.82). Due to the very low number of non-binary participants, it was not possible to calculate agreement for this number. Agreement was also good for whether recruitment information was reported (Fleiss' κ = 0.62), and whether gender was discussed and/or analysed (Fleiss' κ = 0.66). The four coders then annotated the remaining 614 papers, randomly allocated so that each coder annotated approximately the same number of papers. In cases where some participant data were reported as being excluded from analyses we aimed only

[6]https://humanrobotinteraction.org/

to count and report on those participants whose data was actually used in analysis. However, this was not always possible; where some papers reported the demographics of participants *recruited*, then indicated they dropped data from e.g. n = 10, others reported the demographics of participants whose data was actually *analysed*. We hence flagged and captured the dropping of participants as additional participant data (see below).

Offenwanger et al's paper provides full detail on the guidelines underpinning the data collection process [59]. Here, we summarise key decisions and assumptions important for interpreting the results presented in Section 5.

3.1.1 The Binary Assumption. The dataset only contains 'raw' information from the papers, meaning that we did not try to interpret it at this stage. For example, if a paper referred to "20 participants (10 women)", we reported 20 total participants, of which 10 women and 10 of unknown gender, and that the paper utilised a binary gender assumption. However, for calculating gender metrics we interpreted this to mean 10 men and 10 women.

3.1.2 The Othering of Non-Binary Participants. Where papers utilised an *other* category in reporting participant gender, we assumed (both during data collection and in participant counts) that these participants were non-binary individuals, on the basis that they did not identify with binary male or female terms.

3.2 Classification of (Additional) Participant Data

In order to analyse participant sources as Offenwanger et. al did [59], we tagged all text that contained additional information about study participants. This typically included items such as age, nationality, familiarity with robots, etc. Notably, we paid particular attention to reporting (or lack thereof) of participant recruitment method, given Offenwanger at al.'s suggestion that gender bias in research participation might stem from recruitment method. We undertook post-hoc classification on these items, creating a set of labels which were applied to papers based on the collected data. The full list of labels and criteria for them to be applied can be found in Supplementary Materials Table I.

3.3 Classification of Gender Analyses and Discussion

We also identified papers that conducted some form of analysis of participant gender. Three coders (the three authors most familiar with HRI research) individually annotated these papers, which were then explicitly discussed in order to reach an agreement on what the analyses pertained to. This was required due to ambiguity in the way we saw such analyses being reported. As a result, we classified papers that had a clear research question or hypothesis related to gender as 'main gender discussion', and further separated these into papers that analysed the relevant results qualitatively and/or quantitatively. The remaining papers, which treated gender as 'confound', 'controlled' for gender when conducting statistical analysis, or did some post-hoc analysis, were labelled as 'confound'.

3.4 Automatically Identifying Sub-Topics of HRI

In order to classify the papers by HRI sub-topic, we followed the method used by Offenwanger et al. [59], applying probabilistic topic modelling [12] to our 684 papers using the MALLET library [35]. Based on coherence and perplexity measures, we determined that a suitable set of topics would most likely lie in the range of 8 to 20 topics. Topic sets in these counts were visualized as word clouds, and independently labelled, based both on the wordcloud results as well as which papers were associated with each topic, by three of the authors who have been part of the HRI community for several years. The results were discussed to select a final list of 15 (non-exclusive) HRI sub-topics

(see Table 2, original wordclouds are provided in Supplementary Materials Figure 1(a)). We excluded two topics ('User Study' and 'Data, Systems' respectively) from final analyses due to their high paper count and high generality (all but 2 papers tagged with one of these labels were also tagged with at least one other). The paper count for each topic can be found in Table 2. The distribution of sub-topics for each year of the conference can be seen in Supplementary Materials Figure 5

4 A SURVEY OF HRI RESEARCHERS

We conducted a short online survey targeting HRI researchers in order to collect additional information about who is conducting HRI research. Specifically, we asked respondents about: their experience in HRI (years in the field) and with the HRI conference (number of first and co-authored papers accepted for publication); gender; educational background; current field; location; research topic (selecting from our identified sub-topics of HRI with an option to add more topics); and their research practices, specifically which participant information they typically collect and why (with response items designed based on what we saw most frequently during our systematic review). The survey questions are provided in Supplementary Materials Table II.

The primary motivation for undertaking this survey is to explore any correlations between (diversity in) researcher identity and participant diversity across different sub-topics, motivated by the initial evidence for such presented in [59]. Asking researchers to select sub-topics from the same list we identified from the systematic review allowed us to correlate trends in that sub-topic with author identity (gender, education, specialism) and practices. Whilst this might have resulted in a somewhat 'looser' connection between our survey respondents and the dataset from our systematic review, we preferred this approach to any attempt to post-hoc identify e.g. the gender of specific authors of publications in our dataset (as done e.g. by [41]).

For analyses pertaining to correlations in researcher versus participant diversity, we utilise survey data only for those respondents who identified as having authored or co-authored one or more publications published at the HRI conference. For other data presented in the supplementary materials, e.g. regarding data collection practices and educational background, we utilise data from all respondents.

Participants were recruited via international robotics and HRI mailing lists, social media advertisement and through the authors' research networks and were not compensated for their participation. A total of 113 researchers completed the survey (see Section 5.2 for their demographics). Figure 1 shows participants' years of experience working in HRI broken down across those who did and did not report publishing work at the HRI conference. Notably, 73 (65%) of the respondents reported publishing at the conference, and it is the data from these respondents that we use when investigating the correlation between author and participant diversity. This survey should be interpreted with some key caveats. Even though we distributed it across various HRI news channels, the people who took part might not have been an accurate representation of who conducted HRI research (and/or published at the conference) across these last 15 years. Also, it is possible that the people we recruited might have been already interested in issues of diversity, which would be reflected in their research practice.

5 FINDINGS

5.1 Who is *Taking Part* in HRI Research? (RQ 1,2)

Concerning RQ1 on HRI participation, we find (similar to [59]) that gender reporting is increasing (Figure 2). Also, our analysis reveals that, of the papers that report gender, men have made up the majority of HRI research participants to date, with a small but consistent gap between men and women (Figure 3 left) that shows no indication of changing (Figure 3 right). Across all HRI

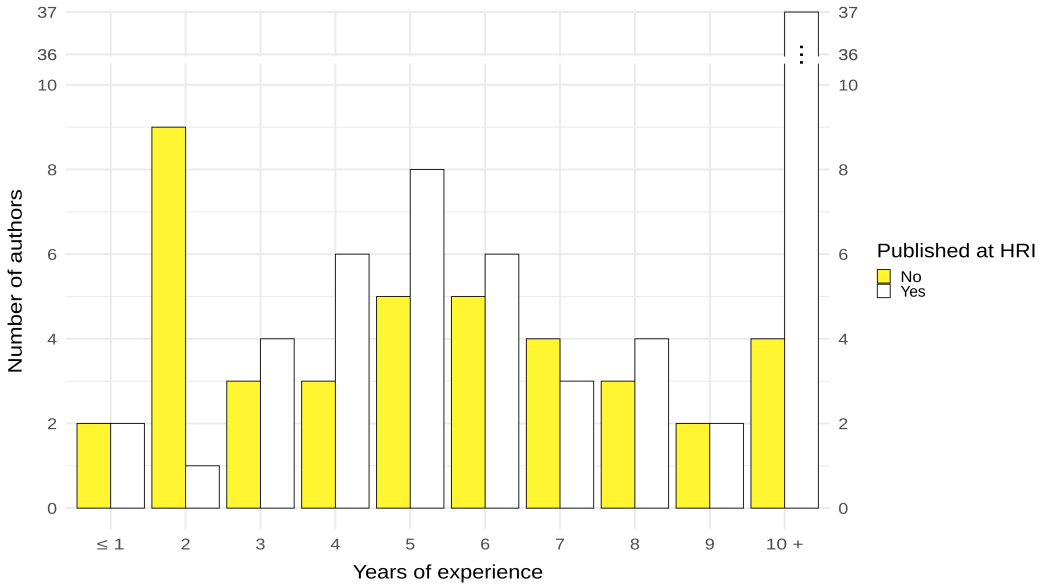

Fig. 1. Variation in years of experience across survey respondents who have/have not published at the HRI conference.

conference papers between 2006 and 2021, average DER was -0.085, and has fluctuated around an average of -0.066 since 2010.

Mirroring Offenwanger et al.'s findings from HCI [59], non-binary participants represent only a tiny proportion of participants in HRI research studies (making them barely visible on Figure 4). Only in 2020 and 2021 have more than two papers at the conference reported studies including non-binary participants, with these participants representing 0.18% and 0.46% [7] of the total participants described in those years, respectively. From the current data it is not possible to know the extent to which this under-representation might be caused by authors providing only binary answer options in gender demographics questions. Further, only 6 of the 17 papers across 2020 and 2021 reporting studies with non-binary participants avoided *othering* by specifically utilising gender terms such as *non-binary* rather than simply listing anyone who didn't identify with pre-defined items under an 'other' category, deviating from published best practices for the inclusive capture and reporting of participant gender [71].

Concerning RQ2a on variation in gender bias patterns across different research topics, Figure 5 (left) shows (by use of the DER measure) that participant gender diversity varies across sub-topics of HRI, with a Kruskal-Wallis test demonstrating some of these differences to be significant ($\chi^2(13) = 36.639$, $p < 0.001$). Specifically, we found that the subtopic of "control and teleoperation" had significantly lower DER compared to "groups", "gaze and attention", "gender and anthropomorphism", and "design and design studies". The mean and standard deviation of DER for each subtopic, and post-hoc pairwise comparisons using Dunn tests, are provided in Table 2 and Supplementary Materials Table III.

Concerning RQ2b regarding the impact that recruitment method might have on gender diversity in research participants, we found that only three recruitment practices were referred to in five or

[7] we note that 0.46% might be considered somewhat representative based on US population statistics based (see https://williamsinstitute.law.ucla.edu/publications/nonbinary-lgbtq-adults-us/)

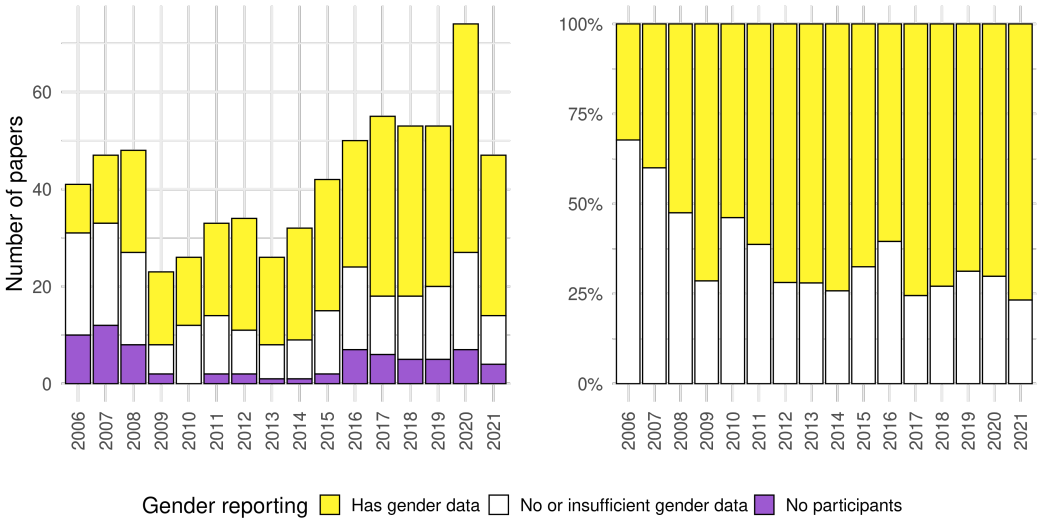

Fig. 2. Gender reporting by year for each HRI conference from 2006-2021. Left: raw number of papers at each year with/without participants and gender reporting. The papers marked as having "no participants" refer to papers that did not report a user evaluation or user study; as can be seen, the majority of papers published at the HRI conference report some sort of study with human subjects. Right: the percentage of papers (with participants) reporting participant gender. Gender reporting appears to have increased but is not ubiquitous, and there is no indication of a continuing trend.

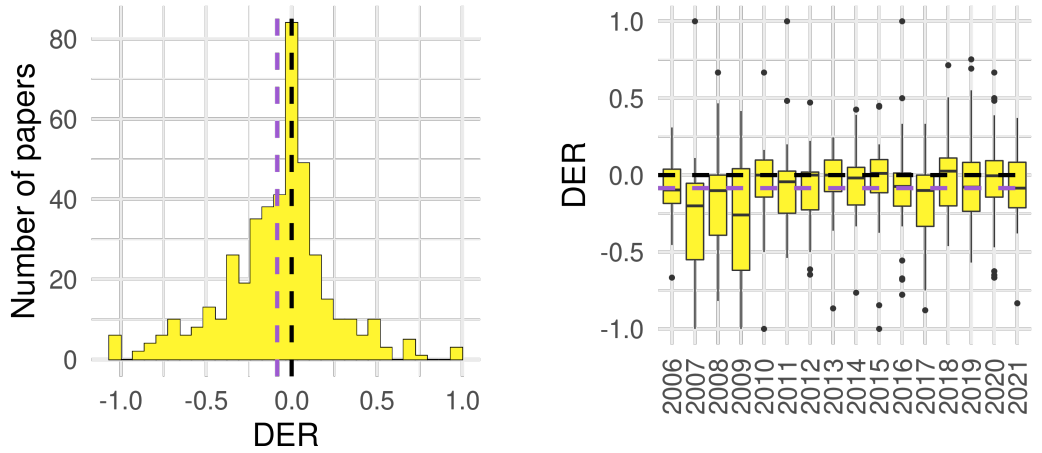

Fig. 3. Left: histogram of paper DER of men and women [59] based on the 446 papers from which DER could be calculated. DER = 0.0 is marked by the black dashed line, whereas the purple dashed line marks the mean DER across all papers (−0.08). Right: DER of papers published at each year of the conference.

more papers across multiple years of the conference (see Table I in the supplementary materials for our coding schema). These were *community* and *local institution* recruitment — which appear to have remained consistently common across all years of the conference — and *crowdsourcing* — the majority of which specifically refers to *Amazon Mechanical Turk (MTurk)*, which has increasingly

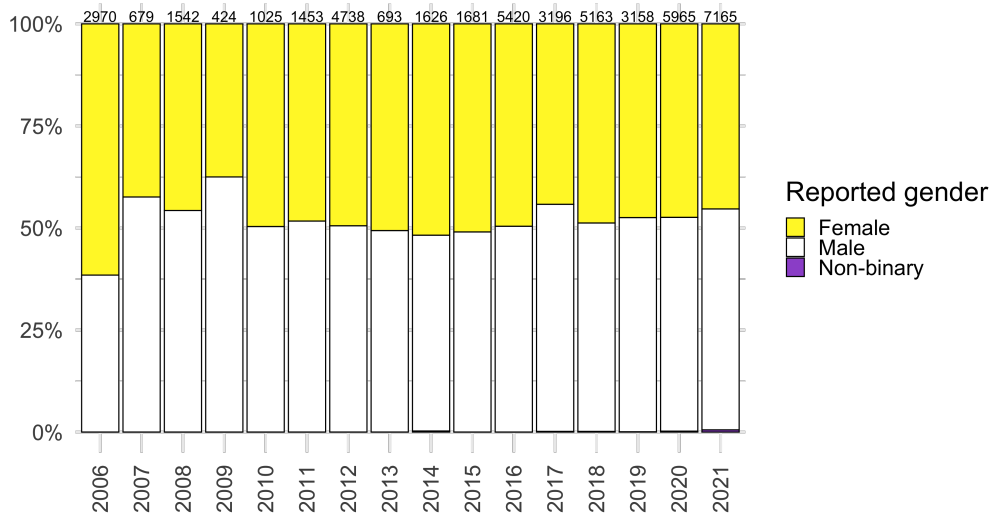

Fig. 4. Total (normalised) count of participants of different genders across all papers which reported participant gender data. The number of people reported as non-binary was zero for all years except 2014 (4 people), 2017 (5 people), 2018 (8 people), 2019 (2 people), 2020 (13 people), and 2021 (39 people).

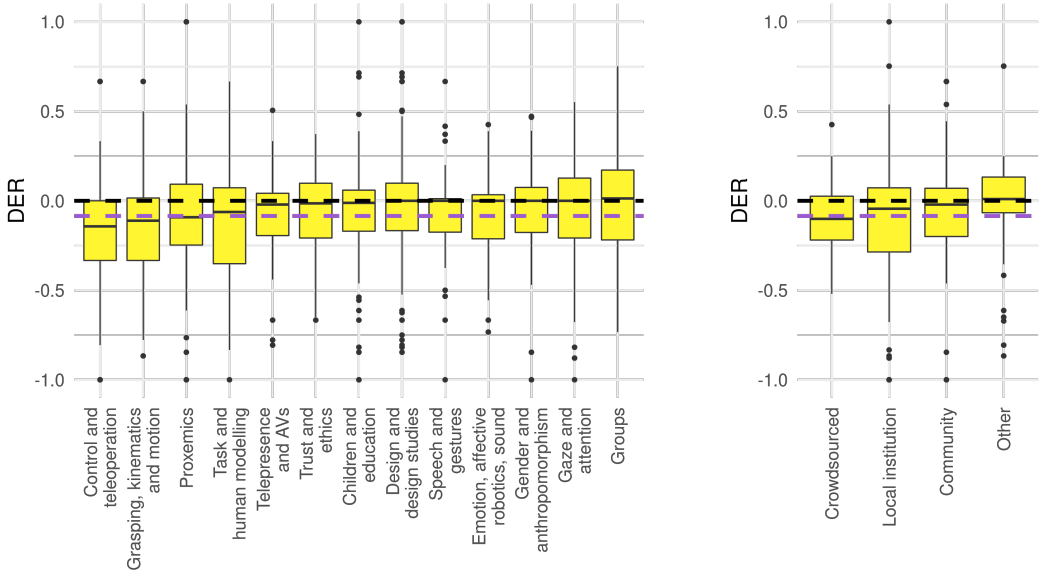

Fig. 5. Left: DER by sub-topic of HRI based on our dataset of HRI papers from 2005-2021. DER = 0.0 is marked by the black dashed line, whereas the purple dashed line marks the mean DER across all papers (−0.08); the *Groups* subtopic was the only topic to have a non-negative DER, with a mean DER of 0.001. Right: DER across the most common recruitment practices consistently referred across multiple papers and multiple years of the conference.

Table 1. Gender distribution of survey respondents divided by whether they published at least one paper at an HRI conference.

Gender	Published at HRI	Not published at HRI
Man	38	23
Woman	33	15
Non-binary	0	2
Genderfluid	1	0
I prefer not to say	2	1

appeared since 2015. Specifically, of the 65 papers that used crowdsourcing methods, 56 (86%) used MTurk.

The COVID-19 pandemic likely made in-person user studies infeasible for many researchers in 2020 and 2021; however, given that papers published at HRI 2020 were submitted in autumn 2019, this alone does not explain the increase in MTurk studies between 2019 and 2020. Similarly, if the pandemic was specifically responsible for a (temporary) increase in the use of MTurk for recruitment, then a larger increase than demonstrated might also have been expected between HRI 2020 and HRI 2021. As such, it seems reasonable to conclude there is a steady increase in the use of MTurk for recruiting HRI research participants, regardless of the pandemic.

Figure 5 (right) shows DER across the different recruitment types reported in papers, which was significantly different (Kruskal-Wallis test, $\chi^2(3) = 12.24$, $p < 0.01$). We replicate Offenwanger et al's finding that crowdsourcing seems to represent a potential source of gender bias within participants, because papers citing its use have significantly lower DER than papers utilising other recruitment methods (post-hoc Dunn test, $Z = -2.92, p = .01$. The full table of post-hoc pairwise comparisons can be found in Supplementary Materials Table IV).

5.2 Who is *Doing* HRI Research? (RQ3)

Of the 113 survey respondents, 61 identified as men, 48 as women, 2 as non-binary, 1 as genderfluid; 3 researchers opted not to share their gender (note that participants could select more than one gender descriptor, in line with best practice guidelines [71]). Respondents' gender distribution is shown in Table 1. Respondents reported working across 22 countries, distributed as follows: USA (41), Sweden (14), the Netherlands (10), Germany (9), the UK (8), Canada (5), Austria (3), Portugal (3), Switzerland (2), Spain (2), New Zealand (2), Japan (2), Italy (2), Denmark (2), France (1), Israel (1), India (1), Singapore (1), China (1), Turkey (1), Belgium (1) and Australia (1). As previously mentioned in Section 4, 73 respondents (65%) reported publishing at least one first or co-authored paper at an HRI conference. Figure 6 reports the number of HRI researchers working on the subtopics we previously identified during the systematic review. Due to the low number of respondents who identified as non-binary, genderfluid, or preferred not to disclose this information, we removed them from this figure as it would potentially constitute identifiable information.

In the supplementary materials, we provide additional results regarding the number of sub-topics participated reported working on (Supplementary Materials Figure 2) and their different educational backgrounds versus current fields (Supplementary Materials Figure 3). Supplementary Materials Figure 3 in particular represents a study in what more intersectional consideration of diversity and representation issues in HRI might 'look like', as one might consider how educational background and gender differently impact on movement and/or belonging (building on [83]) within different sub-topics of research.

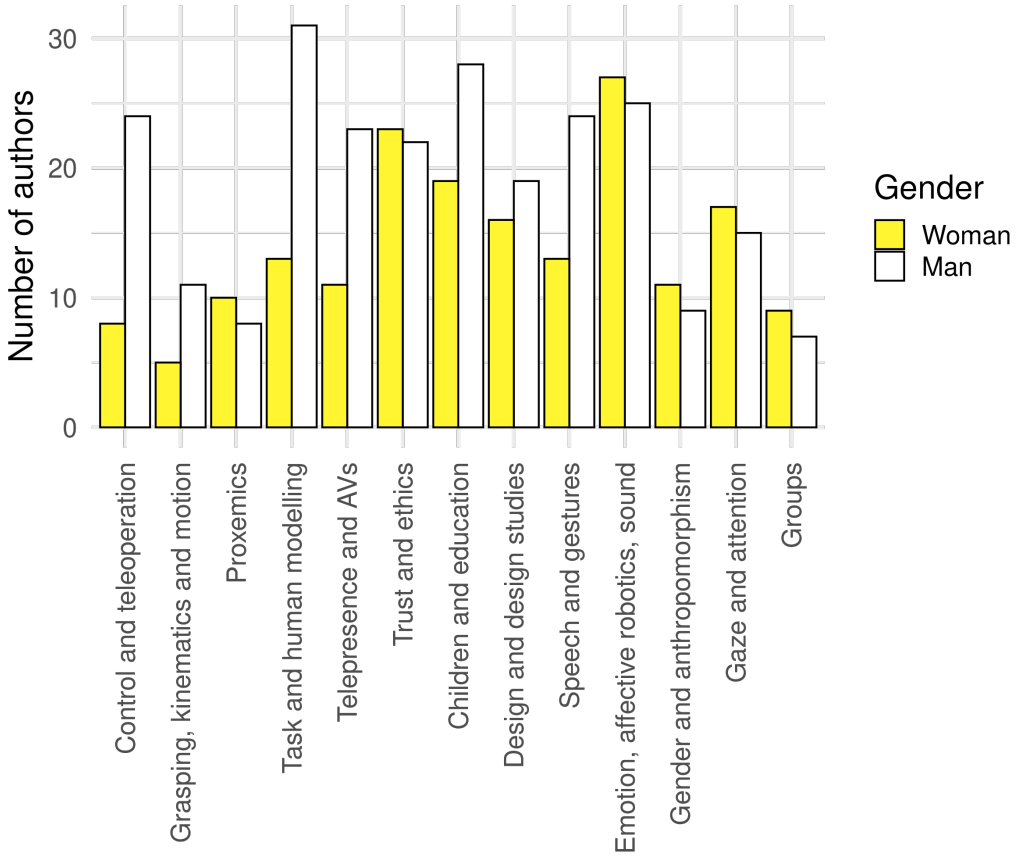

Fig. 6. Number of people working in our identified sub-topics of HRI, divided by gender (note that respondents could identify more than one area, and also add additional responses in free text form; three respondents only used free text answers for this question and are hence not represented here). Due to the low number of respondents who identified as non-binary, genderfluid, or preferred not to disclose this information, we removed them from this figure as it would potentially constitute identifiable information.

Regarding researcher diversity, we computed the GDI (Gender Diversity Index) proposed in [41] (Section 2.3), with data from the author survey. Based on a count of self-identified men, women, non-binary, genderfluid respondents (note that we did not include people that preferred not to say), we calculated a GDI of 0.58. In comparison, the average GDI of Affective Computing and Intelligent Interaction (ACII) authors from [41] was 0.85 which would imply a higher diversity in ACII compared to HRI. However, this comparison needs to be interpreted with some caveats. While we calculated the GDI based on several self-reported gender categories, whereby respondents could also enter free text, [41] inferred binary gender categories from authors' names. As mentioned in Section 2.3, smaller categories are penalised in the Pielou index; thus, even though our current GDI appears to be faring worse than the one computed for the ACII conference, it might be a by-product of trying to evade binary assumptions. Calculating the GDI of our respondents while only considering men and women, the result becomes 0.99, which is supposedly much more diverse.

Table 2. Topic number, author generated label, paper count and mean and standard deviation in DER for each of the sub-topics of HRI generated via probabilistic topic modelling. Note the *User Study* and *Data, Systems* topics were identified as general topics not useful for analysis are also included for reference.

N	Topic Label	Paper Count	M DER	SD
14	Groups	66	.00	0.30
7	Design and Design Studies	241	-0.01	0.32
13	Gaze and Attention	123	-0.04	0.30
12	Gender and Anthropomorphism	134	-0.05	0.24
11	Emotion; Affective Robotics; Sound	82	-0.06	0.25
6	Children and Education	117	-0.07	0.33
8	Trust and Ethics	55	-0.07	0.25
2	User Study (Generic Topic)	567	-0.07	0.31
10	Speech and Gestures	113	-0.08	0.26
3	Telepresence and AVs	62	-0.08	0.27
5	Proxemics	126	-0.08	0.34
1	Task and Human Modelling	147	-0.12	0.35
9	Grasping, Kinematics, Motion	71	-0.14	0.34
0	Data, Systems (Generic Topic)	409	-0.14	0.33
4	Control and Teleoperation	140	-0.21	0.33

Furthermore, the Shannon indices (before normalisation) of our respondents are 0.81 and 0.69 when including 4 and 2 genders respectively. At a first glance it might seem that such unweighted indices might be better suited for variables with many categories, however, the lack of an upper boundary makes this number prone to misinterpretation and hijacking. This strengthens our belief that current diversity measures do not properly account for social diversity, and that novel measures are sorely needed. Since it is not possible to capture diversity in general with any one simple measure or model [49, 60], the work of developing such measures for HRI needs to start by identifying what diversity should mean in HRI and based on that design the measures [39]. This might result in several measures to be used in parallel. When using these measures, they should always be presented in conjunction with an appropriate qualitative interpretation grounded in the underlying definition of diversity.

5.3 *Who* is Doing *What* in HRI Research and *How?* (RQ4,5)

Addressing RQ4, Figure 7 shows the relationship between average paper DER (based on our dataset of HRI papers 2006-2021) and average author DER (based on our HRI researcher survey considering only those n = 73 respondents who identified as having published at the conference). There is a moderate correlation between these two values ($r = 0.55, p = .054$), suggesting a connection between gender diversity of authors within a sub-topic and gender diversity of the participants taking part in user studies relating to that topic.

We might expect these relationships to further interact with researcher educational background and specialism, given known gender trends across e.g. psychology versus robotics and AI [59], which ought to be considered in line with a fully intersectional approach to understanding these trends. More diverse research teams might be more aware of under-representation issues, and therefore strive to attract a more representative sample. Based on the responses of our survey, the most commonly reported backgrounds and current fields among HRI researchers are computer science, engineering, social science, and psychology, with the first two being biased towards men

Table 3. The DER for each background and current field, based on the responses to our survey. Each respondents impact to each field is weighted based on numbers of fields selected; a respondent selecting two fields as their background is counted as 0.5 in both selected fields. This is the reason for the number of participants in each category (N) not being an integer. Mathematics and Physical Sciences have been omitted from the table due to the low number of respondent in those groups (≈ 1 in background, < 1 in current field).

	Background		Current Field	
Field	N	DER	N	DER
Computer Science	38.7	-0.09	44.7	-0.15
Engineering	31.8	-0.39	19.3	-0.42
Social Sciences	6.4	0.40	13.8	0.14
Psychology	11.3	0.09	14.2	0.07
Humanities	4.3	0.41	2.4	0.13
Self-described	14.2	-0.06	14.1	-0.04

and the latter two being biased towards women (see Table 3). There is, however, a fair bit of mobility between researcher background and (current) researcher specialism/field (see Supplementary Materials Figure 3).

Several survey respondents highlighted their aspiration for representative and diverse sampling, both for scientific rigour and for more meaningful impact. A smaller number of respondents specifically highlighted that race/ethnicity needs more attention in HRI — again echoing similar critique published by HCI researchers [18]. Only 10 papers did such reporting; most of them are recent (the first being in 2010 and 5 being published since 2018), which might indicate an improvement due to raised awareness. With more participant information required for a richer understanding of context, one respondent proposed that: *"Maybe the conferences could encourage reporting of detailed participant information if it somehow did not count against the page limits and there was a standard list of demographic questions that people could work from."*

On RQ5, the vast majority of papers reporting gender did not conduct gender-based analysis (the number of papers per year conducting such analyses can be seen in Supplementary Materials Figure 6). Of the 95 papers that did conduct gender analyses (representing 21% of the total papers which reported participant gender, Figure 8A), the majority (64 papers, representing 67%) treated it as a confound, with only 31 being identified as having a gender-related research question or hypothesis (Figure 8B). Results on the impact of gender are mixed (Figure 8C. This figure is further broken down by sub-topics in Supplementary Materials Figure 7). Concerning the extent to which gender is of *explicit* experimental interest in HRI overall, papers with an explicit research question or hypotheses represent 7.4% of all papers describing a user study in our systematic review. This is in line with additional findings from our researcher survey regarding respondents data collection and reporting, presented in the Supplementary Materials.

6 DISCUSSION AND IMPLICATIONS

6.1 Gender Bias in HRI Participation

Our findings suggest HRI researcher mirrors HCI research with regards to gender bias in research participation, as we document an over-representation of men alongside very few non-binary participants to date. We similarly found these bias patterns to vary across (i) recruitment method, in particular replicating a concerning trend of decreasing women taking part in (increasingly popular)

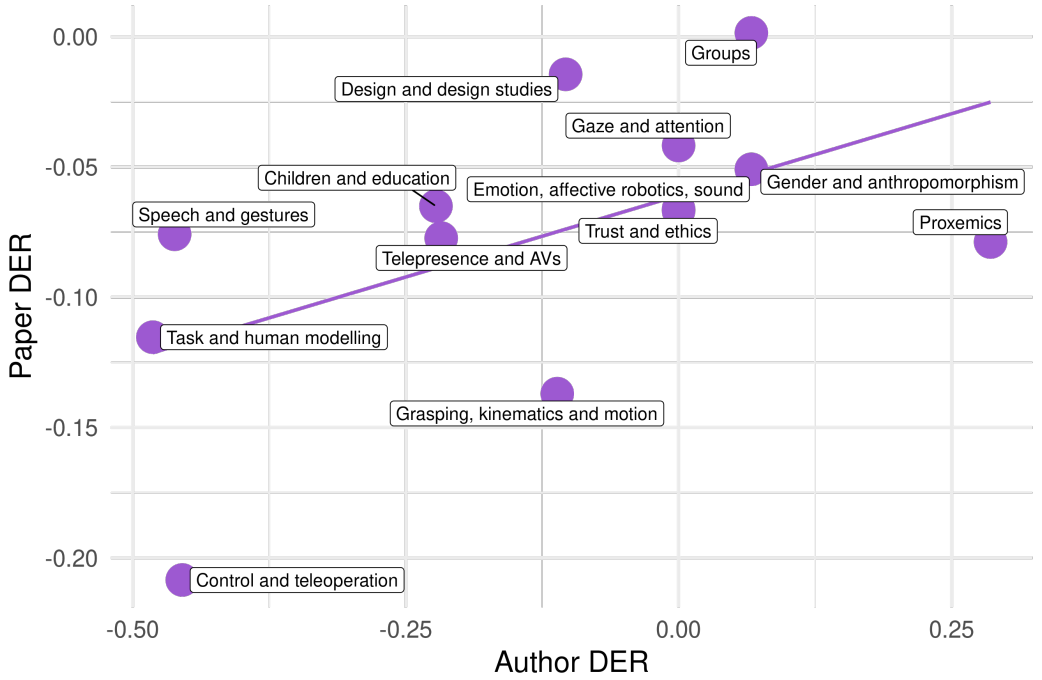

Fig. 7. Average paper DER plotted against author DER.

crowdsourced studies and (ii) HRI sub-topic, which in turn likely links to diversity of researchers (in terms of gender but also educational background and specialism) working on those sub-topics.

6.1.1 The Pros and Cons of Crowdsourcing. Crowdsourcing was the only recruitment method to be associated with a significant decrease in DER (similar to findings for user studies undertaken by the CHI community [59]). This is concerning given the popularity of the the MTurk platform in particular, which has also been condemned for poor ethical practices regarding e.g. workers' compensation and rights [36]. Other crowdsourcing websites such as Prolific [8] have been posited as a potentially more ethical alternative for conducting high quality online research [46], and specifically provide gender screening tools that allow for targeted recruitment of participants that might be used to increase participant diversity. Of course, such efforts are subject to (1) the gender representation of workers on those platforms and (2) the use of such screeners e.g. to explicitly engage with (rather than exclude-by-design) non-binary participants to avoid thus reinforcing their current under-representation in research. Prolific in particular "went viral" on TikTok in August 2021 leading to *'30,000 new participant signups to Prolific, which skewed heavily towards female participants in their 20s'* [9] and have also recently updated their gender screening options to reflect that participant sign-ups are asked *"What gender are you currently? We will ask about your sex later"* with options whereby e.g. *Woman* explicitly includes *Trans Female / Trans Woman.* At the time of writing, the site reports a total of 24,969 men, 40,883 women and 1,777 non-binary participants as being active on the site in the past 90 days. Prolific's other screening options offer an excellent example of why, in practice, gender must be considered intersectionally, as we have alluded to

[8]https://prolific.co/
[9]https://blog.prolific.co/we-recently-went-viral-on-tiktok-heres-what-we-learned/

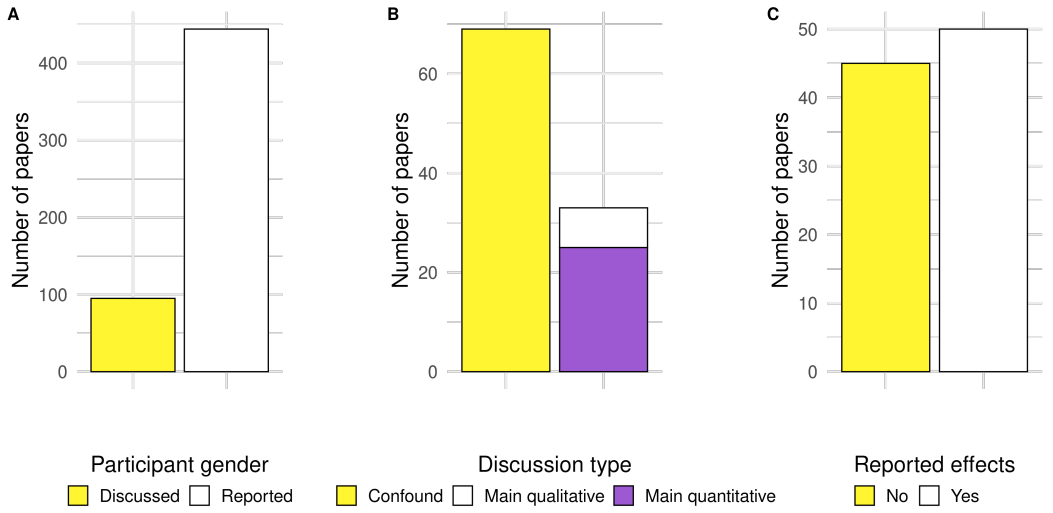

Fig. 8. A: the number of papers in our dataset which report and/or conduct some sort of analysis relating to participant gender. B: how participant gender was treated by those papers reporting a gender based analysis (main hypothesis versus confound). C: the number of papers which reported gender effects versus those which either specifically reported no effect or, identified including gender in their analysis but did not report any related gender specific effects.

throughout this article. Whilst there are many more women active on Prolific than men, applying a screener to find participants whose job involves supervising 10+ subordinates identifies 1,406 men, 1,635 women and 39 non-binary participants who meet this criteria; the relative number of men compared to women thus being hugely different to the overall proportion of men versus women active on the site.

6.1.2 Variations Across HRI Sub-Topics and Researcher Diversity. Again mirroring trends seen in CHI, we found that participant gender representation significantly varied across sub-topics of HRI research, with papers classified as *control and teleoperation*, *grasping, kinematics and motion*, *proxemics* and/or *task and human modelling* typically having greater over-representation of men. In the Introduction, we identified both moral and practical reasons for maximising diversity in user testing, and we encourage researchers in these fields to reflect on whether (a) papers they're citing and/or generalising from might actually present biased results and (b) whether particular recruitment practices could be limiting their participant diversity.

Except for the *proxemics* topic, these topics were self-selected by more men than women in our survey of HRI researchers, in-line with our finding that researcher DER somewhat correlates with paper DER. This might partly be explained by variations in sense of belonging among individuals of different identities in different topics [83], but would seemingly point to another benefit of diversity and collaboration in research teams, something we suggest HRI is well-poised to offer given gender and research specialism/fields documented e.g. in Figures 2 and 3.

6.2 The Reporting of Participant Gender and Conducting of Gender Analyses

Whilst the majority of user study papers we analysed reported participants' gender, it is perhaps surprising that this was not universal (even in papers from the most recent editions of the conference) given e.g. the APA guidelines we refer to in the Introduction. Whilst we did not replicate the entirety

of Offenwanger et al.'s detailed analysis of gender/sex language being used to report participants [59], we did identify that many papers engaged in the *othering* of non-binary participants. Therefore, we encourage HRI researchers to integrate current best practices for the inclusive collection and analyses of participant gender into their work [71]. We also note that more widespread utilisation of these practices would improve the clarity of gender reporting.

Whilst the majority of papers we reviewed reported gender to some extent, and our researcher survey also appeared to suggest the main reason for collecting participant gender was for reporting purposes, it was surprisingly difficult for us to consistently extract quantitative data about exactly who had taken part in the studies described. Similarly, we do not know from our current data to what extent the low number of non-binary participants stems from authors failing to provide participants with a non-binary option in their gender demographics question, or due to the explicit recruitment of men and women only (although this was made explicit in some works e.g. [45, 91]. We therefore suggest that authors always report the percentage of participants who selected each gender option presented, even when that percentage amounts to 0, and make it clear whether (and why) any gender screening was applied at the point of participant recruitment or data analyses.

6.2.1 A Divisive Topic in HRI?. Gender analysis was a key topic of discussion at the 2022 HRI conference, spotlighted in Friederike Eyssel's keynote as she called for more of such (linking to her work with Tannenbaum et al [85]) and again at the co-hosted DEI workshop[10], in which we also presented interim results of our systematic literature review. We took that opportunity to discuss the divergent reviews we had received on this work to date, primarily focused on comments from the HRI Conference AC committee members who ultimately chose to reject an earlier version of this manuscript submitted for publication at the 2022 conference. Despite good reviews from the assigned reviewers, committee members were concerned that our work could be seen to be encouraging the very gender-based analysis that Eyssel seemed to be calling for. We highlight this as further evidence that we, the HRI community, have divergent views on the what, whys and hows on accounting for gender in HRI design and evaluation, and want to share those reviewers' concerns that, whilst more gender analysis might be desirable, it should of course be well motivated to avoid post-hoc rationalisation of findings that could propagate gender stereotypes. A classic example of such rationalisation (as pointed to by one of our reviewers) is the phrenological idea that women must be less intelligent than men because they have smaller heads [42]. Indeed, other social psychology research has posited that gender stereotypes emerge specifically in order to rationalise distribution of the sexes into social roles [38]; and the phrenology example demonstrates how quantitative 'evidence' can be used to further legitimise such ideas. Whilst the question of when gender analyses is (not) appropriate is hugely important then, it is not one we set out to answer with this manuscript, and must surely be answered in an interdisciplinary and intersectional manner. Again, we point to Burtscher and Spiel's article on the topic of gender sensitivity in HCI as an excellent and accessible starting point in this direction [18].

We should take care to consider who we are inviting to take part in our research, and, in cases where gender analysis is deemed appropriate, be cognisant that the statistical methods most commonly used in HRI to date fundamentally prohibit the inclusion of the very small number of non-binary participants. Studies of the general need to be complemented by studies of the particular. Quantitative methods might support some element of generalisation (although feminist approaches would warn us that universality is neither feasible nor necessarily desirable [6]), but to be able to say something about the general, the unusual is often clustered or ignored, making it even less visible [27]. Qualitative approaches instead focus on highlighting the particular, allowing the reader to identify what and how it is relevant for their situation [61]. The transferability of a study is thus

[10]https://sites.google.com/view/dei-hri-2022/home

dependent on rich descriptions of the context and situation. For instance, our data on the relations between HRI researchers' background versus current research field and topic (see Supplementary Materials Figure 3) is not large enough to draw many strong genereliseable conclusions, but it is rich enough to identify some patterns for further study.

6.3 Summary of Practical Suggestions Moving Forward

Here we pull together and summarise some starting recommendations on how we, the HRI research community, might be able to improve on participant (gender) diversity and representation moving forwards. Notably these recommendations cover actions at different levels (e.g. from individual researchers/reviewers through to conference/journal organisation committees), and are intended to provide a starting point for further discussions and initiatives. These recommendations further build on existing guidelines [71] and recommendations [18] coming from our neighbouring field of HCI.

6.3.1 Collecting, Reporting and Analysing Participant Gender. At the individual level, we suggest researchers can:

(1) follow existing, published guidelines for inclusive gender capture and reporting [71]
(2) report how gender demographics questions were asked, i.e. whether participants were asked about sex, gender and/or what answer options they were able to elect from (including options that were not ultimately selected by any participants)
(3) be clear in the motivations for any gender analyses/hypothesised effects, and consider pre-registering such analyses to avoid falling in to post-hoc rationalisations
(4) recognise that gender is only one identity characteristic pertinent to research participant experience, likely to intersect with other facets of identity, being aware hence of the limitations of their (gender) analyses [18]

At the community level, we might consider how conferences and journals can provide (minimum, enforced and outside of page limit) requirements for extent and style of demographic information (cf. e.g. the information on ethical recruitment required for submissions to the Interaction Design and Children Conference[11]).

At both levels, these efforts could also draw from (and contribute to) initiatives like the IEEE's work in progress on a standard for recommended practice regarding human subjects research in human-robot interaction[12].

6.3.2 Qualitative Data Collection and Alternative Approaches to Thinking About Gender. The overwhelming majority of papers we identified as undertaking some sort of gender analysis were quantitative. Whether in gender analyses or in attempts to quantify diversity, numbers cannot speak for themselves, they need to be understood in the appropriate context. This is particularly important when complex phenomena such as diversity is condensed into individual indices. The benefit of such numbers is the overview and comparisons it facilitates, but the cost is, by necessity, disregarding the complexity. When using indices like that it is therefore crucial to complement them with a thorough exposition and discussion of underlying assumptions, definitions, and context. This could also be an excellent point where quantitative measures can interface in a constructive way with qualitative results and reflections. On an individual level then, researchers might (re-)consider when quantitative gender is appropriate, at the very least being transparent and clear in communicating any possible limitations of such when sharing results, as well as if/how they might use qualitative data collection/analysis to complement such analyses. Similarly, educators can draw

[11] https://idc.acm.org/2022/ethics-inclusion-and-accessibility/
[12] https://standards.ieee.org/ieee/3108/10710/#Working

attention to (and validate) the use of qualitative methods in HRI (e.g. utilising material from the recent HRI textbook [8]) and ensure their syllabi highlight the limitations of quantitative analyses. An obvious benefit of qualitative research is the reduced emphasis on needing large sample sizes to support quantitative, categorical analysis, making it easier to include (rather than inherently excluding) participants of less common gender identities. When discussing issues regarding identity in computing, educators should also be aware of how stereotypes in the field can affect the sense of belonging among the students, leading to implicit exclusion of some subgroups [83].

At the community level, we can also look to ensure qualitative and/or other alternative methodologies examining the impacts of gender (for example) on HRI are not disadvantaged at the review stage, simply due reviewers being more comfortable and familiar with quantitative analyses. Again, this is something that the interdisciplinarity of HRI ought to be well-placed to support, provided that researchers with the related expertise are represented in reviewer pools and program committees. We should recognise the research opportunities that exist in bringing non-dominant perspectives to HRI, and create spaces where discussion of these ideas can be led by appropriate experts. Workshops like the 2021 and 2022 *Gendering Robots* workshop co-located with the ROMAN conference[13] offer such a vehicle for further encouraging interdisciplinarity, the sharing of best practices and the mainstreaming ideas from outside of traditional computer science/engineering within HRI. We (at both the individual and community levels) must be open to new and emerging methods of analysis from these and other related fields. We point readers again, as a first example, to the previously mentioned recent Nature article calling for context over categories in social psychological theory [23]. The discussion of social constructionism, assemblage theory and dynamic systems, alongside how these might be operationalised in examining social phenomena, might provide some inspiration for alternative ways to approach HRI research.

6.3.3 Participant Recruitment. At the individual level, researchers might reflect on their choice of participant recruitment method, in particular around crowdsourcing and the risks/benefits of gender screening tools provided by some crowdsourcing platforms. Burtscher and Spiel provide some recommendations specifically pertinent to the recruitment and study of (gender) minority participants, e.g. recrtuiment via peer groups, respect of gender self-identity and creation of non-judgemental safe spaces [18]. At an institutional level, support could be provided to increase accessibility through e.g. facilitating the proper re-numeration of participant labour and in ensuring accessibility of experimental spaces. If engaging in more qualitative data collection as discussed above, researchers can also better include participants of e.g. non-binary gender identity, but we caution against researchers trying to target recruit, examine and report on the experiences of persons with a particular identity in which they do not share – see discussions pertaining to collaboration, rather than extraction of community knowledge, e.g. in Data Feminism [27].

6.3.4 Leveraging and Expanding Diversity in HRI. As a research field, HRI is diverse in terms of contributing disciplines, which (currently) seemingly also brings with it some element of gender diversity, as we have e.g. more men than women coming from computer science and more women than men coming from the social sciences and psychology (see e.g. Table 3). We should strive to extend and capitalise on such diversity within our field, as more diverse research teams are more likely, in turn, to recruit more diverse participants. Such efforts must also include those who are LGBTQ2IA* or of a minority race identity, and the social and/or professional networking events typically employed at most academic conferences could be adapted specifically towards creating such connections and trying to foster greater sense of belonging [83]. When creating such structures, it is important to scrutinise who are actually encouraged and allowed to participate and

[13]https://sites.google.com/view/ro-man-genr-workshop/home

to belong, and who are implicitly or explicitly excluded. Initiatives such as Queer in AI[14] and Black in Robotics[15] are working towards improved inclusiveness in the fields by compiling resources, building networks, developing guidelines, etc. that we can take inspiration from in this regard.

At the individual level, researchers might therefore consider who they are (not) collaborating with, and, in the case of faculty or lab managers, who they are recruiting into their groups. This must be done cautiously however in order to avoid *tokenism*, particularly given the potential to create teams which replicate current gender trends (e.g. hiring one woman researcher with a psychology background in an effort to simultaneously increase gender and field diversity). To avoid this, faculty, group leaders and others in positions of power in particular might consider taking on professional development activities concerned with identity in computing to better understand diversity, equity and inclusion (DEI) issues in both our research and in our institutions (see e.g. The Cultural Competence in Computing (3C) Fellows Program[16]).

At the community level, we might also consider how we can raise awareness and understanding of these issues via e.g. choice of conference keynote speakers, conference-hosted tutorials akin to those seen at the CHI conference and workshops dedicated to the topic, the DEI workshop at HRI 2022 providing a great starting point. Building on our discussion of diversity metrics (and their limitations), we also identify a (community level) need for work towards identifying what aspects of diversity are particularly relevant for HRI, and how those would most appropriately be operationalised into comparable measures. This would not circumvent the need for grounding and discussing the results in each case they are used (see again the discussion on quantitative plus qualitative data in the previous subsection), but would facilitate interpretations and comparisons within the context of HRI. In addition, it is important monitor how the community members feel in relation to the field to prevent people only being included on paper, and make sure that there is an actual sense of belonging. In particular, acknowledging that participation and belonging is not equally obvious for everyone [83].

7 CONCLUSION

In this paper, we have outlined our definition of gender bias (Section 1.2), described our collection of both participant (Section 3) and researcher gender and background information (Section 4), then using the results to to address our research questions (Section 5) and provide discussion points including a summary of practical suggestions for improved practice moving forward (Section 6).

We found in response to RQ1 that there is some evidence of gender bias in the HRI research field (section 5.1), and that this bias in representation follows similar patterns to what was found in HCI by Offenwanger et al. [59]. These repeated patterns also extend to RQ2 about how the bias varies, where, again like HCI, we see gender representation differing between subtopics of HRI, and across different recruitment practices, specifically in the use of crowdsourcing, although we identify available tools that might be used to counteract this in section 6.1.1. Turning to RQ3 and the question of who is doing what in HRI, our analysis clearly points to the need for further work on diversity measures (section 5.2), but we also show interesting trends in gender and field mobility, which merit further investigation (Figure 3). This is all the more important in light of RQ4, where we show that there is some evidence for a link between researcher identity and participant diversity (section 5.3), but as researcher gender, background, and field are are tied together, further analysis with a focus on intersectionality will be required to make progress on understanding this complex phenomenon.

[14]https://www.queerinai.com/
[15]https://blackinrobotics.org/
[16]https://identity.cs.duke.edu/fellows.html

In our reflection on gender analysis in HRI, we find in response to RQ5 that it is a minority of papers which conduct some form of gender based analysis. Those that do such analysis typically treat gender as a confound within their main statistical analysis (section 5.3). It is difficult to say whether or not this is a problem in and of itself, but there is a clear need for further discussion on this topic within the HRI research community (see the discussion in section 6.2.1). As a starting point, we identify some of the key arguments for/against gender based analysis, and note the potential for more qualitative treatments of gender to simultaneously support inclusion of non-binary individuals and better contextual understanding of gender effects.

This research contributes a comprehensive data set of reported participant gender in HRI research, as well as additional information about author gender and background, the joint analysis of which show interesting trends, and highlights areas in need of future work. We highlight a clear need for further discussions of gender practices within HRI research and suggest the practical suggestions we provide in our discussion summary (section 6.3) are a concrete step in that direction. The findings reported in this article are only a subset of the rich amount of information that can be extracted from our dataset. While in the current paper we have focused on gender representation and analyses, we encourage researchers to use it to uncover more trends and patterns in the HRI field.

ACKNOWLEDGEMENTS

We first want to acknowledge those authors who came before us in identifying issues around gender in HCI – these works have significantly informed this manuscript, but also our approaches to HRI research more broadly. We wish to specifically thank Dongwook Yoon and Julia Bullard for their work on the conceptualisation and development of the gender data schema and extraction method that we build on [59]; also Minsuk Chang, Alan Milligan, and Austin Kobayashi for their input and work on initial version of the MAGDA tool [59].

We would like to thank all of the HRI researchers who engaged with our survey. We additionally want to thank those reviewers who provided constructive feedback on an earlier version of this work, and reviewers for/attendees of the DEI Workshop held at HRI 2022 for further discussion — the manuscript is improved greatly as a result and we hope we have done justice to both the positive and negative critique we have received. All figures were generated with a colour palette based on the non-binary pride flag, created by Joel Le Forestier (https://joelleforestier.com/#pridepalettes). This work was partially funded by the Digital Futures Research Centre.

REFERENCES

[1] 2020. What is gender? What is sex? https://cihr-irsc.gc.ca/e/48642.html Accessed: Dec 09, 2021.

[2] 2021. Könsidentitet och könsuttryck. https://www.1177.se/liv--halsa/konsidentitet-och-sexuell-laggning/konsidentitet-och-konsuttryck/ Accessed: Dec 22, 2021.

[3] Veronica Ahlqvist, Winnie Birberg, Magnus Lagerholm, Anders Sundin, Carl Sundström, and Lisbeth Söderqvist. 2020. *Uppföljning av Vetenskapsrådets implementering av köns- och genusperspektiv i forskningens innehåll.* Memorandum Dnr 3.3-2018-05687. Swedish Research Council. https://www.vr.se/analys/rapporter/vara-rapporter/2020-02-06-uppfoljning-av-vetenskapsradets-implementering-av-kons--och-genusperspektiv-i-forskningens-innehall.html

[4] Majed Al Zayer, Isayas B. Adhanom, Paul MacNeilage, and Eelke Folmer. 2019. The Effect of Field-of-View Restriction on Sex Bias in VR Sickness and Spatial Navigation Performance. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300584

[5] Thomas Arnold and Matthias Scheutz. 2018. Observing Robot Touch in Context: How Does Touch and Attitude Affect Perceptions of a Robot's Social Qualities?. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (Chicago, IL, USA) *(HRI '18)*. Association for Computing Machinery, New York, NY, USA, 352–360. https://doi.org/10.1145/3171221.3171263

[6] Shaowen Bardzell. 2010. Feminist HCI: taking stock and outlining an agenda for design. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1301–1310.

[7] Louise Barkhuus and Jennifer A. Rode. 2007. From Mice to Men - 24 Years of Evaluation in CHI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 1 pages. https://doi.org/10.1145/1240624.2180963

[8] Christoph Bartneck, Tony Belpaeme, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Šabanović. 2020. *Human-robot interaction: An introduction.* Cambridge University Press.

[9] Paul Baxter, James Kennedy, Emmanuel Senft, Severin Lemaignan, and Tony Belpaeme. 2016. From characterising three years of HRI to methodology and reporting recommendations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 391–398.

[10] Nancy E Betz and Louise F Fitzgerald. 1987. *The career psychology of women.* Academic Press.

[11] Vijay S Bhagat. 2018. Women authorship of scholarly publications in STEMM: Authorship puzzle. *Feminist Research* 2, 2 (2018), 66–76.

[12] David M Blei. 2012. Probabilistic topic models. *Commun. ACM* 55, 4 (2012), 77–84.

[13] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) *(NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 4356–4364.

[14] Hannah Louise Bradwell, Katie Jane Edwards, Rhona Winnington, Serge Thill, and Ray B Jones. 2019. Companion robots for older people: importance of user-centred design demonstrated through observations and focus groups comparing preferences of older people and roboticists in South West England. *BMJ open* 9, 9 (2019), e032468.

[15] De'Aira Bryant, Jason Borenstein, and Ayanna Howard. 2020. Why Should We Gender? The Effect of Robot Gendering and Occupational Stereotypes on Human Trust and Perceived Competency. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20)*. Association for Computing Machinery, New York, NY, USA, 13–21. https://doi.org/10.1145/3319502.3374778

[16] M. Burnett, R. Counts, R. Lawrence, and H. Hanson. 2017. Gender HCl and microsoft: Highlights from a longitudinal study. In *2017 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, Raleigh, NC, USA, 139–143. https://doi.org/10.1109/VLHCC.2017.8103461

[17] Margaret Burnett, Simone Stumpf, Jamie Macbeth, Stephann Makri, Laura Beckwith, Irwin Kwan, Anicia Peters, and William Jernigan. 2016. GenderMag: A method for evaluating software's gender inclusiveness. *Interacting with Computers* 28, 6 (2016), 760–787.

[18] Sabrina Burtscher and Katta Spiel. 2020. " But where would I even start?" developing (gender) sensitivity in HCI research and practice. In *Proceedings of the Conference on Mensch und Computer*. 431–441.

[19] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 981–992. https://doi.org/10.1145/2858036.2858498

[20] Juan Pablo Chaverri, Adrián Vega, Kryscia Ramírez-Benavides, Ariel Mora, and Luis Guerrero. 2021. Exploratory Analysis of Research Publications on Robotics in Costa Rica Main Public Universities. In *International Conference on Applied Human Factors and Ergonomics*. Springer, 95–102.

[21] Vijay Chidambaram, Yueh-Hsuan Chiang, and Bilge Mutlu. 2012. Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. 293–300.

[22] Meia Chita-Tegmark, Theresa Law, Nicholas Rabb, and Matthias Scheutz. 2021. Can You Trust Your Trust Measure?. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 92–100.

[23] Mina Cikara, Joel E Martinez, and Neil A Lewis. 2022. Moving beyond social categories by incorporating context in social psychological theory. *Nature Reviews Psychology* (2022), 1–13.

[24] Julia R Cordero, Thomas R Groechel, and Maja J Mataric. [n. d.]. A Review and Recommendations on Reporting Recruitment and Compensation Information in HRI Research Papers. ([n. d.]).

[25] C. R. Crowelly, M. Villanoy, M. Scheutzz, and P. Schermerhornz. 2009. Gendered voice and robot entities: Perceptions and reactions of male and female subjects. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 3735–3741. https://doi.org/10.1109/IROS.2009.5354204 ISSN: 2153-0866.

[26] Kerstin Dautenhahn, Michael Walters, Sarah Woods, Kheng Lee Koay, Chrystopher L Nehaniv, A Sisbot, Rachid Alami, and Thierry Siméon. 2006. How may I serve you? A robot companion approaching a seated person in a helping context. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. 172–179.

[27] Catherine D'ignazio and Lauren F Klein. 2020. *Data feminism.* MIT press.

[28] Virginia Dignum, Melanie Penagos, Klara Pigmans, and Steven Vosloo. 2020. *Policy guidance on AI for children (draft).* Technical Report. Technical Report. UNICEF. https://www. unicef. org/globalinsight/reports ....

[29] Friederike Eyssel. 2022. What's Social about Social Robots? A Psychological Perspective. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction* (Sapporo, Hokkaido, Japan) *(HRI '22)*. IEEE Press, 2.

[30] Julia Fink, Séverin Lemaignan, Pierre Dillenbourg, Philippe Rétornaz, Florian Vaussard, Alain Berthoud, Francesco Mondada, Florian Wille, and Karmen Franinović. 2014. Which robot behavior can motivate children to tidy up their toys? Design and Evaluation of" Ranger". In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. 439–446.

[31] Jodi Forlizzi. 2007. How robotic products become social products: an ethnographic study of cleaning in the home. In *2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 129–136.

[32] Jodi Forlizzi and Carl DiSalvo. 2006. Service Robots in the Domestic Environment: A Study of the Roomba Vacuum in the Home. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction* (Salt Lake City, Utah, USA) *(HRI '06)*. Association for Computing Machinery, New York, NY, USA, 258–265. https://doi.org/10.1145/1121241.1121286

[33] Ana Freire, Lorenzo Porcardo, and Emilia Gómez. 2021. Measuring Diversity of Artificial Intelligence Conferences. In *Artificial Intelligence Diversity, Belonging, Equity, and Inclusion*. PMLR, 39–50.

[34] Aimi S Ghazali, Jaap Ham, Emilia I Barakova, and Panos Markopoulos. 2018. Effects of robot facial characteristics and gender in persuasive human-robot interaction. *Frontiers in Robotics and AI* 5 (2018), 73.

[35] Shawn Graham, Scott Weingart, and Ian Milligan. 2012. *Getting started with topic modeling and MALLET*. Technical Report. The Editorial Board of the Programming Historian.

[36] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P Bigham. 2018. A data-driven analysis of workers' earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–14.

[37] Marie Hicks. 2019. Hacking the Cis-tem. *IEEE Annals of the History of Computing* 41, 1 (2019), 20–33.

[38] Curt Hoffman and Nancy Hurst. 1990. Gender stereotypes: Perception or rationalization? *Journal of personality and social psychology* 58, 2 (1990), 197.

[39] Sönke Hoffmann and Andreas Hoffmann. 2008. Is there a "true" diversity? *Ecological Economics* 65, 2 (2008), 213–215.

[40] Wei Huang. 2015. When HCI Meets HRI: the intersection and distinction. *Virginia Polytechnic Institute and State University, January* (2015).

[41] Isabelle Hupont, Songül Tolan, Ana Freire, Lorenzo Porcaro, Sara Estevez, and Emilia Gómez. 2021. How diverse is the ACII community? Analysing gender, geographical and business diversity of Affective Computing research. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–8.

[42] Janet Shibley Hyde. 1990. Meta-analysis and the psychology of gender differences. *Signs: Journal of Women in Culture and Society* 16, 1 (1990), 55–73.

[43] Refugees Immigration and Citizenship Canada. 2019. Canadians can now identify as gender "X" on their passports. https://www.canada.ca/en/immigration-refugees-citizenship/news/notices/gender-x-documents.html Accessed: Dec 09, 2021.

[44] Bahar Irfan, James Kennedy, Séverin Lemaignan, Fotios Papadopoulos, Emmanuel Senft, and Tony Belpaeme. 2018. Social psychology and human-robot interaction: An uneasy marriage. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 13–20.

[45] 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*. 559–567.

[46] Patrik Jonell, Taras Kucherenko, Ilaria Torre, and Jonas Beskow. 2020. Can we trust online crowdworkers? Comparing online and offline participants in a preference test of virtual agents. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–8.

[47] Lou Jost. 2006. Entropy and diversity. *Oikos* 113, 2 (2006), 363–375.

[48] Lou Jost. 2007. Partitioning diversity into independent alpha and beta components. *Ecology* 88, 10 (2007), 2427–2439.

[49] Kenneth Junge. 1994. Diversity of ideas about diversity measurement. *Scandinavian Journal of Psychology* 35, 1 (1994), 16–26.

[50] Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction* 2, CSCW (2018), 1–22.

[51] Erik Lagerstedt and Serge Thill. 2020. Benchmarks for evaluating human-robot interaction: lessons learned from human-animal interactions. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 137–143. https://doi.org/10.1109/RO-MAN47096.2020.9223347

[52] Nichola Lubold, Erin Walker, and Heather Pon-Barry. 2016. Effects of voice-adaptation and social dialogue on perceptions of a robotic learning companion. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 255–262.

[53] C. Mendez, A. Sarma, and M. Burnett. 2018. Gender in Open Source Software: What the Tools Tell. In *2018 IEEE/ACM 1st International Workshop on Gender Equality in Software Engineering (GE)*. IEEE, Gothenburg, Sweden, 21–24.

[54] Marek P Michalowski, Selma Sabanovic, and Hideki Kozima. 2007. A dancing robot for rhythmic social interaction. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*. 89–96.

[55] Corinne A Moss-Racusin, John F Dovidio, Victoria L Brescoll, Mark J Graham, and Jo Handelsman. 2012. Science faculty's subtle gender biases favor male students. *Proceedings of the national academy of sciences* 109, 41 (2012), 16474–16479.

[56] Justin Munafo, Meg Diedrick, and Thomas A Stoffregen. 2017. The virtual reality head-mounted display Oculus Rift induces motion sickness and is sexist in its effects. *Experimental brain research* 235, 3 (2017), 889–901.

[57] Mathias Wullum Nielsen, Jens Peter Andersen, Londa Schiebinger, and Jesper W Schneider. 2017. One and a half million medical papers reveal a link between author gender and attention to gender and sex analysis. *Nature human behaviour* 1, 11 (2017), 791–796.

[58] Tatsuya Nomura. 2016. Robots and Gender. *Gender and the Genome* 1, 1 (Dec. 2016), 18–25. https://doi.org/10.1089/gg.2016.29002.nom Publisher: Mary Ann Liebert, Inc., publishers.

[59] Anna Offenwanger, Alan John Milligan, Minsuk Chang, Julia Bullard, and Dongwook Yoon. 2021. Diagnosing bias in the gender representation of HCI research participants: how it happens and where we are. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.

[60] Scott E. Page. 2011. *Diversity and Complexity*. Princeton University Press.

[61] Michael Quinn Patton. 2015. *Qualitative evaluation and research methods* (4 ed.). SAGE Publications, inc, Chapter 81, 710–721.

[62] Caroline Criado Perez. 2019. *Invisible women: Exposing data bias in a world designed for men*. Random House, New York City, United States.

[63] Aaron Powers. 2008. FEATURE What robotics can learn from HCI. *Interactions* 15, 2 (2008), 67–69.

[64] Julia Rosén, Jessica Lindblom, and Erik Billing. 2021. Reporting of Ethical Conduct in Human-Robot Interaction Research. In *International Conference on Applied Human Factors and Ergonomics*. Springer, 87–94.

[65] Matthew B Ross, Britta M Glennon, Raviv Murciano-Goroff, Enrico G Berkes, Bruce A Weinberg, and Julia I Lane. 2022. Women are credited less in science than men. *Nature* 608, 7921 (2022), 135–145.

[66] María Teresa Ruiz-Cantero, Carmen Vives-Cases, Lucía Artazcoz, Ana Delgado, Maria del Mar García Calvente, Consuelo Miqueo, Isabel Montero, Rocío Ortiz, Elena Ronda, Isabel Ruiz, et al. 2007. A framework to analyse gender bias in epidemiological research. *Journal of Epidemiology & Community Health* 61, Suppl 2 (2007), ii46–ii53.

[67] Selma Šabanović. 2010. Robots in society, society in robots. *International Journal of Social Robotics* 2, 4 (2010), 439–450.

[68] Maha Salem, Micheline Ziadee, and Majd Sakr. 2014. Marhaba, How May i Help You? Effects of Politeness and Culture on Robot Acceptance and Anthropomorphization. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction* (Bielefeld, Germany) *(HRI '14)*. Association for Computing Machinery, New York, NY, USA, 74–81. https://doi.org/10.1145/2559636.2559683

[69] Paul Schermerhorn, Matthias Scheutz, and Charles R Crowell. 2008. Robot social presence and gender: Do females view robots differently than males?. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*. 263–270.

[70] Morgan Klaus Scheuerman, Jacob M. Paul, and Jed R. Brubaker. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 144 (Nov. 2019), 33 pages. https://doi.org/10.1145/3359246

[71] Morgan Klaus Scheuerman, Katta Spiel, Oliver L Haimson, Foad Hamidi, and Stacy M Branham. 2020. HCI guidelines for gender equity and inclusivity. *UMBC Faculty Collection* (2020).

[72] Matthias Scheutz and Thomas Arnold. 2016. Are we ready for sex robots?. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 351–358.

[73] Ari Schlesinger, W. Keith Edwards, and Rebecca E. Grinter. 2017. Intersectional HCI: Engaging Identity through Gender, Race, and Class. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5412–5427. https://doi.org/10.1145/3025453.3025766

[74] Mariah L Schrum, Michael Johnson, Muyleng Ghuy, and Matthew C Gombolay. 2020. Four years in review: Statistical practices of likert scales in human-robot interaction studies. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 43–52.

[75] SFS 1972:119. [n. d.]. Lag om fastställande av könstillhörighet i vissa fall. https://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/lag-1972119-om-faststallande-av_sfs-1972-119

[76] SFS 2018:162. [n. d.]. Lag om statlig ersättning till personer som har fått ändrad könstillhörighet fastställd i vissa fall. https://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/lag-2018162-om-statlig-ersattning-till_sfs-2018-162

[77] Vivek K. Singh, Mary Chayko, Raj Inamdar, and Diana Floegel. 2020. Female librarians and male computer programmers? Gender bias in occupational images on digital media platforms. *Journal of the Association for Information Science and Technology* 71, 11 (2020), 1281–1294. https://doi.org/10.1002/asi.24335 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24335

[78] Gabriel Skantze. 2017. Predicting and regulating participation equality in human-robot conversations: Effects of age and gender. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*. IEEE, 196–204.

[79] SOU 2014:91. 2014. Juridiskt kön och medicinsk könskorrigering. https://www.regeringen.se/rattsliga-dokument/statens-offentliga-utredningar/2015/01/sou-201491-/

[80] SOU 2017:92. 2017. Transpersoner i Sverige. https://www.regeringen.se/rattsliga-dokument/statens-offentliga-utredningar/2017/11/sou-201792/

[81] Ian F Spellerberg and Peter J Fedor. 2003. A tribute to Claude Shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon–Wiener'Index. *Global Ecology and Biogeography* 12, 3 (2003), 177–179.

[82] Katta Spiel. 2021. " Why are they all obsessed with Gender?"—(Non) binary Navigations through Technological Infrastructures. In *Designing Interactive Systems Conference 2021*. 478–494.

[83] Jane G Stout and Heather M Wright. 2016. Lesbian, gay, bisexual, transgender, and queer students' sense of belonging in computing: An Intersectional approach. *Computing in Science & Engineering* 18, 3 (2016), 24–30.

[84] Yolande Strengers and Jenny Kennedy. 2020. *The smart wife: Why Siri, Alexa, and other smart home devices need a feminist reboot.* MIT Press.

[85] Cara Tannenbaum, Robert P Ellis, Friederike Eyssel, James Zou, and Londa Schiebinger. 2019. Sex and gender analysis improves science and engineering. *Nature* 575, 7781 (2019), 137–146.

[86] Beverly Daniel Tatum. 2017. *Why are all the Black kids sitting together in the cafeteria?: And other conversations about race.* Hachette UK.

[87] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. *International AAAI Conference on Weblogs and Social Media* 9 (01 2015), 454–463.

[88] Emily Wang, Constantine Lignos, Ashish Vatsal, and Brian Scassellati. 2006. Effects of Head Movement on Perceptions of Humanoid Robot Behavior. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction* (Salt Lake City, Utah, USA) *(HRI '06)*. Association for Computing Machinery, New York, NY, USA, 180–185. https://doi.org/10.1145/1121241.1121273

[89] Yan Wang and James E. Young. 2014. Beyond "Pink" and "Blue": Gendered Attitudes towards Robots in Society. In *Proceedings of Gender and IT Appropriation. Science and Practice on Dialogue - Forum for Interdisciplinary Exchange* (Siegen, Germany) *(Gender IT '14)*. European Society for Socially Embedded Technologies, Artur-Woll-Haus, Am Eichenhang 50, DEU, 49–59.

[90] Mark West, Rebecca Kraut, and Han Ei Chew. 2019. *I'd blush if I could: closing gender divides in digital skills through education.* Technical Report. https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1

[91] Katie Winkle, Gaspar Isaac Melsión, Donald McMillan, and Iolanda Leite. 2021. Boosting Robot Credibility and Challenging Gender Norms in Responding to Abusive Behaviour: A Case for Feminist Robots. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 29–37.

[92] Holly O Witteman, Michael Hendricks, Sharon Straus, and Cara Tannenbaum. 2019. Are gender gaps due to evaluations of the applicant or the science? A natural experiment at a national funding agency. *The Lancet* 393, 10171 (2019), 531–540.

[93] Hongjun Ye, Haeyoung Jeong, Wenting Zhong, Siddharth Bhatt, Kurtulus Izzetoglu, Hasan Ayaz, and Rajneesh Suri. 2020. The Effect of Anthropomorphization and Gender of a Robot on Human-Robot Interactions. In *Advances in Neuroergonomics and Cognitive Engineering (Advances in Intelligent Systems and Computing)*, Hasan Ayaz (Ed.). Springer International Publishing, Cham, 357–362. https://doi.org/10.1007/978-3-030-20473-0_34