A Systematic Review of Participant Gender in 15 Years of HRI User Studies

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

Erik Lagerstedt*
University of Skövde
Skövde, Sweden
erik.lagerstedt@his.se

Ilaria Torre*
KTH Royal Institute of Technology
Stockholm, Sweden
ilariat@kth.se

Anna Offenwanger*
Université Paris-Saclay
Gif-sur-Yvette, France

***Abstract*—Recent work identified a concerning trend of disproportional gender representation in research participants in Human-Computer Interaction (HCI). Given that many Human Robot Interaction (HRI) methodologies originate in/mirror those seen in HCI, we are investigating the extent to which this trend is replicated in our field through production of a dataset covering participant gender representation, reporting, and analysis in all 684 full papers published at the HRI conference from 2006-2021. This short paper presents a high level overview of some of our most pressing findings, key discussion points, and questions for the HRI community regarding gender in HRI.**

***Index Terms*—Human Robot Interaction, Research Methods, Gender, Diversity**

I. Introduction

Since the start of the IEEE/ACM International Conference on Human-Robot Interaction in 2006, researchers have been concerned with whether user gender might influence human robot interaction (HRI) [1]–[3]. Today, HRI works continue to examine the impact of user gender [4]–[6], robot gendering [7], [8] and if/how these two might interact [9]–[12]. 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 [13], [14], but also through data exclusion [15]–[17], and embedding gender bias directly into machines [17]–[19].

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 [1] in Human-Computer Interaction (HCI) research [20]. 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 continued 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 [21]–[23], Offenwanger et al.'s findings clearly motivate a thorough reflection on *how* we are doing HRI, and perhaps more specifically *who* we are inviting to do it with us and how we are reporting that in our publications.

A. Quantifying Diversity in Research

For this short paper, we focus on replicating the *Distance from Even Representation* (DER) measure developed by Offenwanger et. al. [20]:

$$DER = \frac{women - men}{women + men} \tag{1}$$

ranging from [-1;1], to specifically study the relative representation of men and women, where 0 corresponds to equal representation. Although this metric lacks the ability to handle more than two genders, or intersectional aspects of gender, its simplicity makes interpretations and limitations clearer.

II. The HRI Research Participation Dataset

We annotated the 684 full papers published at the ACM/IEEE HRI conference from 2006 to 2021 (excluding extended abstracts, LBR's, student design competitions, and video submissions).

A. Data Collection Tool and Data Schema

We contacted the authors of Offenwanger et. al. [20], 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.

For brevity in this short paper, we refer readers to Offenwanger et al for full detail on the guidelines underpinning the data collection process [20] but summarise here a few key decisions and assumptions important for interpreting our initial results.

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

[1]We refer readers to [20] for definitions of gender, sex and bias as we utilise these terms in this work

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.

2) The Othering of Non-Binary Participants: Where papers utilised an *other* category in reporting participant gender, we have assumed (both during data collection and in participant counts) that these participants identified as non-binary individuals, on the basis that they did not identify with binary male or female terms. Given that authors may also have included participants who chose not to share their gender within this *other* category, it's possible that we therefore over-estimate the number of non-binary participants who've taken part in research to date.

B. Classification of (Additional) Participant Data

In order to analyse participant sources as Offenwanger et. al did [20], we tagged all text that contained additional information about study participants. This typically included items such as age, nationality, familiarity with robots, etc.

C. Classification of Gender Analyses and Discussion

We also identified papers that conducted some form of analysis of participant gender, or that discussed gender in the paper. 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'.

D. 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. [20], applying probabilistic topic modelling [24] to our 684 papers using the MALLET library [25].

III. Some Key Findings

A. Who is Taking Part in HRI Research?

Our analysis reveals that men have made up the majority of HRI research participants to date, with a small but consistent gap between men and women that shows no indication of changing. Across all HRI conference papers to date, average DER was -0.085 (equivalent to approximately 20% more men than women) and has fluctuated around an average of -0.066 since 2010 (equivalent to approximately 15% more men than women).

Mirroring Offenwanger et al.'s findings from HCI [20], non-binary participants represent only a tiny proportion of participants in HRI research studies. Only in 2020 and 2021 have more than two papers at the conference reported studies including non-binary participants, with these participants representing only 0.18% and 0.46% of the total participants described in those years, respectively.

Notably, only 6 of the 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 [26].

B. Variations Across Sub-Field and Recruitment Method

Figure 1 shows (by use of the DER measure) that participant gender diversity varies across sub-topics of HRI, with a Kruskal-Wallis test demonstrating these differences to be significant ($\chi^2(13) = 36.639$, $p < 0.001$).

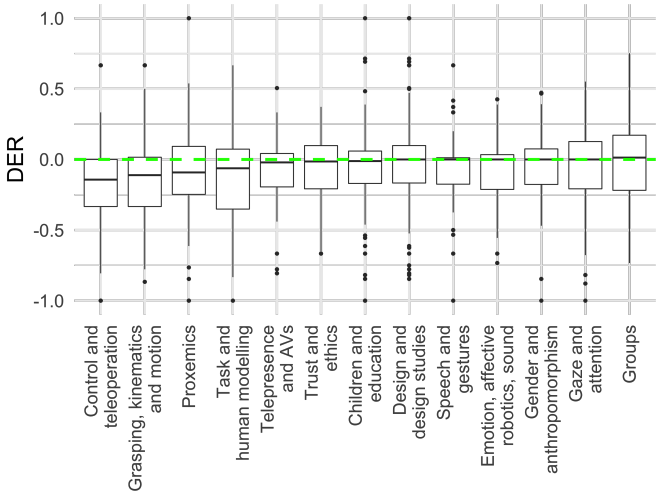


Fig. 1: DER by sub-topic of HRI based on our dataset of HRI papers from 2005-2021. DER = 0.0 is marked by the dashed line. Mean DER across all papers was -0.08, the *Groups* subtopic was the only topic to have a non-negative DER, with a mean DER of 0.001.

We found that only three recruitment methods were referred to in five or more papers across multiple years of the conference (see 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 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 2 shows DER across the different recruitment types reported in papers. We replicate Offenwanger et al's finding that crowdsourcing seems to represent a potential source of gender bias within participants, as papers citing its use have significantly lower DER than papers utilising other recruitment methods (Kruskal-Wallis test, $\chi^2(3) = 12.24$, $p < 0.01$).

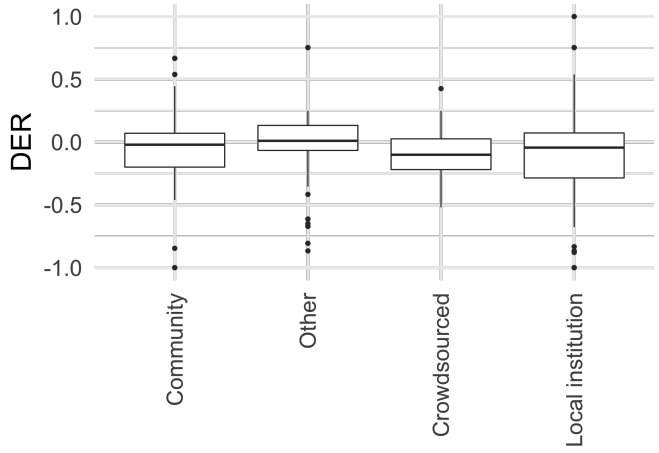


Fig. 2: DER across the most common recruitment methods consistently referred across multiple papers and multiple years of the conference.

C. Gender Analyses in HRI User Studies

The vast majority of papers reporting gender did not conduct gender based analysis, of the 101 papers that did, the majority treated it as a confound. Only 7.4% of all papers describing a user study in our systematic review identified an explicit research question or hypotheses regarding participant gender, and results on the impact of gender thus far are mixed (Figure 3).

IV. Discussion Points and Open Questions for the HRI Community

A. How Do We Improve (and Measure) Diversity in Research Participation?

These initial results point towards HRI sub-field and/or recruitment method (no doubt linked) as influencing research participation. For example, our review identified crowdsourcing as being a potential source of bias. MTurk specifically has also been condemned for poor ethical practices regarding e.g. workers' compensation and rights [27]. However, other crowdsourcing websites such as Prolific [2] have been posited as a potentially more ethical alternative for conducting high quality online research [28], and specifically provide gender screening tools that allow for targeted recruitment of participants by gender that might be used to reduce gender bias. 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'* [3] 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*. Prolific's other screening options offer a fantastic example of the intersectionality we have alluded to throughout this article in practice: researchers can apply a screeners relating to e.g. work, political and religious beliefs, education etc., use of which will no doubt have implications on participant diversity which we ought to be aware of. In addition, we must also be cognisant that targeted recruitment of minority individuals (attempting to increase diversity) often leads to their engaging in disproportionate, often underpaid/valued labour; akin to e.g. the over-burdening of female faculty with service tasks as universities aim for gender balance in key administrative functions [29].

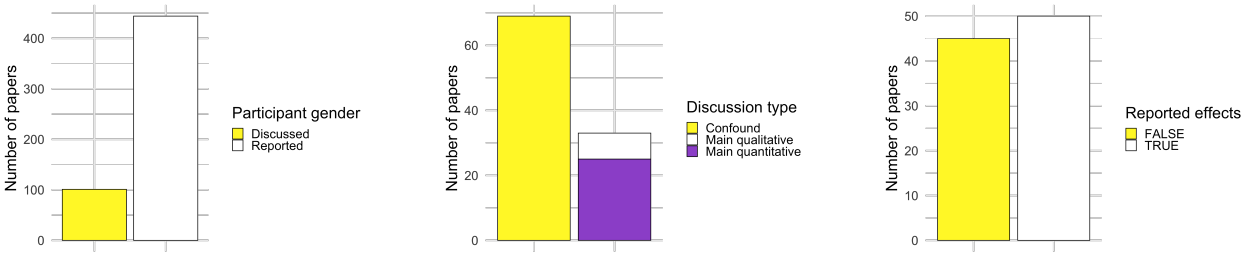
Differences in diversity across sub-fields likely also reflect methodological differences across the various disciplines represented by HRI researchers. Quantitative methods are designed to allow generalization (ideally universally), but to be able to say something about the general, the unusual is often clustered or ignored, making it even less visible [30]. Qualitative approaches instead focus on highlighting specific instances, allowing the reader to identify when and how they can be relevant for their own situations [31]. The transferability of a study is thus dependent on rich descriptions of the context and situation. There is an exciting (and optimistic) question here of whether we can leverage the interdisciplinarity of HRI to improve diversity by sharing best practices cross-discipline.

B. When (not) and How (not) to do Gender Analysis?

Whilst we report on the number of papers which present gender-based analyses, and the proportion of these that treated gender as a main variable of interest versus a potential confound, we want to make it clear that we do not advocate for gender based analysis to be a *default* norm when conducting HRI user studies. A key question we think the community needs to consider (and one for which there is no easy or straightforward answer) is when and how we should (not) be looking for gender effects. On the one hand, neglecting gender differences can lead to the design of systems which disadvantage women and non-binary individuals [15] but unfounded

[2]https://prolific.co/

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



(a) (b) (c)

Fig. 3: (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) and (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.

gender analysis risks the propagation of stereotypes through the post-hoc rationalisation of findings. How do we get past the issue that our most commonly used statistical methods fundamentally push us towards excluding small numbers of e.g. non-binary participants from our analyses? Can recent works e.g. in the areas of Data Feminism [30] and Design Justice [32] help us to answer this question, and better reflect on how we deal with gender (and other intersectional identity traits) in the application, development and testing of HRI?

C. What Should We Report, and How?

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. APA guidelines [4]. As previously noted, we identified a that a number of papers also engaged in the *othering* of non-binary participants. We encourage HRI researchers to integrate current best practices for the inclusive collection and analyses of participant gender into their work [26] but also ask how else we can and should report on study participation. In our data labelling process, it was surprisingly difficult for us to consistently extract quantitative data about exactly who had taken part in the studies described. Further, gender reporting in the papers we analysed left us unable to examine the extent to which the lack of non-binary participants might stem from authors simply failing to provide participants with a non-binary option in their gender demographics question versus targeted and explicit recruitment of men and women only. We therefore suggest that authors always report the number of participants who selected each gender option presented, even when that number is 0, and make it clear whether (and why) any gender screening was applied at the point of participant recruitment or data analyses. Our classification of (additional) participant data also identified a wide range of participant identity traits that may (not) be reported presumably based on their perceived relevance to a paper's research question and/or disciplinary traditions held by the authors. Given the interdisciplinary nature of HRI, we wonder whether the HRI conference might play a greater role in providing reporting guidelines and/or encouraging their use by exluding such reporting from submission page limits. Again, the question of *what* exactly should be reported is one which is not easily answered, and again brings up the notion of intersectionality when considering research participation and its influence on our results.

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

V. Conclusion and Next Steps

We are currently preparing a more detailed manuscript containing additional analysis of our dataset as well as/in combination with results from a survey of HRI researchers. Specifically we look to investigate:

- alternate metrics of diversity (specifically we consider the ecology-inspired *Gender Diversity Index* previously posited as a tool for monitoring diversity in scientific communities [33])
- what (and why) HRI researchers capture and report in the context of user studies
- intersectionality in researcher gender, educational background and sub-field and if/how this correlates with participant diversity

We will release our dataset alongside said manuscript, in the hope that other researchers might investigate other trends in research participation in addition to our gender-focused analysis. Our analysis thus far leads us to the conclusion that user gender currently sits in somewhat of a 'grey area' for researchers, with a lack of clarity surrounding if, when and why gender analyses might be (in)appropriate. Our aim is to produce a summary set of workable, practical suggestions for HRI moving forward, and so we welcome input from the community on the discussion questions raised above.

Acknowledgment

We wish to thank all authors from Offenwanger et al. [20] for providing us with a copy of the MAGDA toolkit, and all of the HRI researchers who engaged with our survey. Figure 3 was generated with a colour palette based on the non-binary pride flag, created by Joel Le Forestier.

References

[1] E. Wang, C. Lignos, A. Vatsal, and B. Scassellati, "Effects of head movement on perceptions of humanoid robot behavior," in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, ser. HRI '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 180–185. [Online]. Available: https://doi.org/10.1145/1121241.1121273

[2] J. Forlizzi and C. DiSalvo, "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*, ser. HRI '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 258–265. [Online]. Available: https://doi.org/10.1145/1121241.1121286

[3] K. Dautenhahn, M. Walters, S. Woods, K. L. Koay, C. L. Nehaniv, A. Sisbot, R. Alami, and T. Siméon, "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*, 2006, pp. 172–179.

[4] P. Schermerhorn, M. Scheutz, and C. R. Crowell, "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*, 2008, pp. 263–270.

[5] G. Skantze, "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, 2017, pp. 196–204.

[6] N. Lubold, E. Walker, and H. Pon-Barry, "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, 2016, pp. 255–262.

[7] D. Bryant, J. Borenstein, and A. Howard, "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*, ser. HRI '20. New York, NY, USA: Association for Computing Machinery, Mar. 2020, pp. 13–21. [Online]. Available: https://doi.org/10.1145/3319502.3374778

[8] H. Ye, H. Jeong, W. Zhong, S. Bhatt, K. Izzetoglu, H. Ayaz, and R. Suri, "The Effect of Anthropomorphization and Gender of a Robot on Human-Robot Interactions," in *Advances in Neuroergonomics and Cognitive Engineering*, ser. Advances in Intelligent Systems and Computing, H. Ayaz, Ed. Cham: Springer International Publishing, 2020, pp. 357–362.

[9] R. B. Jackson, T. Williams, and N. Smith, "Exploring the role of gender in perceptions of robotic noncompliance," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 559–567.

[10] A. S. Ghazali, J. Ham, E. I. Barakova, and P. Markopoulos, "Effects of robot facial characteristics and gender in persuasive human-robot interaction," *Frontiers in Robotics and AI*, vol. 5, p. 73, 2018.

[11] T. Nomura, "Robots and Gender," *Gender and the Genome*, vol. 1, no. 1, pp. 18–25, Dec. 2016, publisher: Mary Ann Liebert, Inc., publishers.

[12] C. R. Crowelly, M. Villanoy, M. Scheutzz, and P. Schermerhornz, "Gendered voice and robot entities: Perceptions and reactions of male and female subjects," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct. 2009, pp. 3735–3741, iSSN: 2153-0866.

[13] C. A. Moss-Racusin, J. F. Dovidio, V. L. Brescoll, M. J. Graham, and J. Handelsman, "Science faculty's subtle gender biases favor male students," *Proceedings of the national academy of sciences*, vol. 109, no. 41, pp. 16 474–16 479, 2012.

[14] M. Burnett, S. Stumpf, J. Macbeth, S. Makri, L. Beckwith, I. Kwan, A. Peters, and W. Jernigan, "Gendermag: A method for evaluating software's gender inclusiveness," *Interacting with Computers*, vol. 28, no. 6, pp. 760–787, 2016.

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

[16] C. Wagner, D. Garcia, M. Jadidi, and M. Strohmaier, "It's a man's wikipedia? assessing gender inequality in an online encyclopedia," *International AAAI Conference on Weblogs and Social Media*, vol. 9, pp. 454–463, 01 2015.

[17] M. K. Scheuerman, J. M. Paul, and J. R. Brubaker, "How computers see gender: An evaluation of gender classification in commercial facial analysis services," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, Nov. 2019. [Online]. Available: https://doi.org/10.1145/3359246

[18] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, "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*, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 4356–4364.

[19] V. K. Singh, M. Chayko, R. Inamdar, and D. Floegel, "Female librarians and male computer programmers? gender bias in occupational images on digital media platforms," *Journal of the Association for Information Science and Technology*, vol. 71, no. 11, pp. 1281–1294, 2020. [Online]. Available: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.24335

[20] A. Offenwanger, A. J. Milligan, M. Chang, J. Bullard, and D. Yoon, "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*, 2021, pp. 1–18.

[21] W. Huang, "When hci meets hri: the intersection and distinction," *Virginia Polytechnic Institute and State University, January*, 2015.

[22] A. Powers, "Feature what robotics can learn from hci," *Interactions*, vol. 15, no. 2, pp. 67–69, 2008.

[23] E. Lagerstedt and S. Thill, "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)*, 2020, pp. 137–143.

[24] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[25] S. Graham, S. Weingart, and I. Milligan, "Getting started with topic modeling and mallet," The Editorial Board of the Programming Historian, Tech. Rep., 2012.

[26] M. K. Scheuerman, K. Spiel, O. L. Haimson, F. Hamidi, and S. M. Branham, "Hci guidelines for gender equity and inclusivity," *UMBC Faculty Collection*, 2020.

[27] K. Hara, A. Adams, K. Milland, S. Savage, C. Callison-Burch, and J. P. Bigham, "A data-driven analysis of workers' earnings on amazon mechanical turk," in *Proceedings of the 2018 CHI conference on human factors in computing systems*, 2018, pp. 1–14.

[28] P. Jonell, T. Kucherenko, I. Torre, and J. Beskow, "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*, 2020, pp. 1–8.

[29] C. M. Guarino and V. M. Borden, "Faculty service loads and gender: Are women taking care of the academic family?" *Research in higher education*, vol. 58, no. 6, pp. 672–694, 2017.

[30] C. D'ignazio and L. F. Klein, *Data feminism*. MIT press, 2020.

[31] M. Q. Patton, *Qualitative evaluation and research methods*, 4th ed. SAGE Publications, inc, 2015, ch. 81, pp. 710–721.

[32] S. Costanza-Chock, *Design Justice: Community-led practices to build the worlds we need*. The MIT Press, 2020.

[33] A. Freire, L. Porcardo, and E. Gómez, "Measuring diversity of artificial intelligence conferences," in *Artificial Intelligence Diversity, Belonging, Equity, and Inclusion*. PMLR, 2021, pp. 39–50.