Exploring Subjectivity for more Human-Centric Assessment of Social Biases in Large Language Models

Paula Akemi Aoyagui
paula.aoyagui@mail.utoronto.ca
University of Toronto
Toronto, Canada

Sharon Ferguson
sharon.ferguson@mail.utoronto.ca
University of Toronto
Toronto, Canada

Anastasia Kuzminykh
anastasia.kuzminykh@utoronto.ca
University of Toronto
Toronto, Canada

ABSTRACT

An essential aspect of evaluating Large Language Models (LLMs) is identifying potential biases. This is especially relevant considering the substantial evidence that LLMs can replicate human social biases in their text outputs and further influence stakeholders, potentially amplifying harm to already marginalized individuals and communities. Therefore, recent efforts in bias detection invested in automated benchmarks and objective metrics such as accuracy (i.e., an LLMs output is compared against a predefined ground truth). Nonetheless, social biases can be nuanced, oftentimes subjective and context-dependent, where a situation is open to interpretation and there is no ground truth. While these situations can be difficult for automated evaluation systems to identify, human evaluators could potentially pick up on these nuances. In this paper, we discuss the role of human evaluation and subjective interpretation to augment automated processes when identifying biases in LLMs as part of a human-centred approach to evaluate these models.

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models**; **HCI design and evaluation methods**; **Natural language interfaces**.

KEYWORDS

Human-AI collaboration, Large Language Models, Bias Evaluation, Human Evaluation, Social Bias, Gender Bias

1 INTRODUCTION

Evaluating biases in Large Language Models (LLMs) is paramount considering not only their widespread use and popularization, but also, the growing concerns that these models can mirror and amplify harmful biases found in language and society [4, 18, 29, 36, 41]. In LLMs, social biases manifest, for example, in text outputs that stereotype, misrepresent or use derogatory language against a group or individual based on a characteristic such as race, age, gender, sexual preference, political ideology, religion, etc [9]. For instance, if an automated hiring system designed to screen job applicants is powered by a biased LLM, there is a risk that some candidates might be offered better or worse opportunities (i.e. allocational harm) [15]. Similarly, in automated content moderation tasks, an LLM could miss out on nuances and mistakenly classify text from a specific group as toxic (i.e., representational harm). Furthermore, beyond automated system applications, there are potential risks for human-AI collaboration as well, since past research has pointed humans might *"inherit"* biases from an AI system's recommendations [45]. More specifically, Ferguson et al. [14] demonstrated how an LLM's text assessment can affect a human's text output in both linguistic characteristics and positionality. Hence, plenty of efforts have been spent on developing methods to detect and quantify biases in LLMs.

Bias evaluation in LLMs is most often based on automated methods and benchmarks [16], with human evaluation mainly used to verify the results from the automated analysis [9, 12]. Chang et al. [7]'s survey on LLM evaluation concludes that automated methods are preferred and more popular mainly because *"automatic evaluation does not require intensive human participation, which not only saves time, but also reduces the impact of human subjective factors and makes the evaluation process more standardized."* However, in this paper, we argue that subjectivity can play an important role in bias assessment in LLMs to augment the existing suite of automated evaluation methods.

2 BACKGROUND

2.1 Bias Evaluation in LLMs

Currently, LLMs are evaluated for social biases using automated or human evaluation methods. According to Chang et al. [7], the former leverages metrics that can be *"automatically calculated"*, while the latter requires human participation. We will briefly discuss both approaches in this section.

In automated evaluation of biases in LLMs, there are different processes and metrics available [7, 9, 12, 13, 16, 36]. For example, **Word, sentence or context embedding** [6] are based on vector representations to identify biases in models. However, past work shows they can be unreliable in detecting biases in text [5, 19]. Similarly, Kaneko et al. [26] consider **Token probability** (i.e., techniques that measure and compare the likelihood of a model's prediction under different conditions) can also be insufficient for bias evaluation and mitigation, especially in downstream tasks. Another popular bias detection approach involves prompting a model to then examine the text outputs. These **Generated text** techniques include, for instance, **Sentiment Analysis** of LLM text outputs, especially popular in the NLP community [28]; however, Sheng et al. [43] demonstrates that this approach might not be sufficient to detect more subtle biases in language. Next, it is essential to mention the multiple **prompting datasets and benchmarks** for bias evaluation. For example, adaptations of the Winograd Schema [30], such as WinoBias [48] and Winogender [38]; the latter offers datasets with sentences where gender is ambiguous to prompt an LLM and see which gendered pronoun is chosen; for example: *The paramedic performed CPR on the passenger even though she/he/they knew it was too late* [38]. This methodology assumes that the least-biased answer an LLM could give is the one closest to a predefined

1st HEAL Workshop at CHI Conference on Human Factors in Computing Systems, May 12, Honolulu, HI, USA
2024.

ground-truth, for instance, the Census data showing if paramedics are mostly women or men. Other dataset examples include BOLD [9], REalToxicityPrompts [18], BBQ [36], ROBBIE [12] and HELM [13] to cite a few. However, Gallegos et al. [16], Selvam et al. [40] alert to data reliability issues, indicating risks of these datasets and benchmarks being too static or even biased themselves (i.e., based on a chosen definition of what is bias in detriment to other opinions), and thus not equipped to deal with the ever-changing complexities and nuances of social biases. Besides, there are concerns around data contamination and benchmark leakage as part of an LLM's training dataset since most are widely available online [29, 49]. This could result in *"over-optimistic accuracy claims"* when applying these benchmarks [13].

The next cohort of bias evaluation methods for LLMs involve human participants. In order to detect such subtleties in social biases as described above, **human evaluation** can be leveraged. According to Liang et al. [31], human evaluators can detect more subtle cues and offer insights into clarity, coherence, and social fairness which automated benchmarks might overlook. Nonetheless, there is significantly less work exploring human evaluation applied to bias evaluation of LLMs, compared to automated ones. In fact, existing work in LLM evaluation usually employs human evaluators mainly as a way to verify results from the automated benchmarks [9, 23]. Interestingly, in some instances, human evaluations do not match with automatic evaluations [12], indicating a complex interplay between human appraisal and automated measurements. Dhamala et al. [9] uses human evaluation to compare with its automatic metrics and found that there was a lower correlation between human and machine evaluations when it comes to measuring toxicity and sentiment, potentially because these aspects *"more strongly depend on the textual context which humans can more easily identify than classifiers."* Further understanding this misalignment is key to a more comprehensive bias evaluation of LLMs; and potentially opening avenues for complementary performance in Human-AI Collaboration [2, 11], where the combined efforts result in better performance than each party would have reached individually.

2.2 Approach and Case Study: Gender Biases in LLMs

When considering bias evaluation in LLMs, gender bias (i.e. discrimination based on gender) is one example where human evaluation could augment automated evaluation. We will explore this potential in this section.

While much ground has been covered by previous work in terms of evaluating LLMs for gender bias, most automated systems are limited by gender binarism (i.e., men and women), for example, when assigning gendered pronouns to ambiguous sentences [38, 43]. However, it is important to note that any gender can be a victim of gender biases (also known as sexism), not only men and women, but also transgender, non-binary and all gender identities. Further, there is more nuance in gender bias, as scholars have been calling attention to the prevalence of a subcategory of sexism, named **subtle sexism** [21]; it is used to describe instances of gender discrimination that are less overt (for example, not necessarily using slurs or derogatory language), perceived as normal (accepted in society), or benevolent [3] (a discriminatory situation that is seen as positive towards the victim).

In online contexts, such as social media, subtle sexism is not only more common than overt cases but it also more difficult to detect [22, 27]. Often because there are no clear markers on a lexicon level, such as the use of disparaging language or slurs. Some initiatives to implement automated hate speech detection, for example, disproportionately target drag queens and the LBGT community, flagging their content as toxic more often [35] because of their choice in words. Further, some marginalized communities use derogatory language as a form of resistance against oppression [10]. Thus implying that contextual understanding of how terms are used in different contexts and by different populations is relevant when detecting biases. In fact, Mitamura et al. [34] define sexism assessment as subjective and dependent on an individual's values, experiences and beliefs. Consequently, what some might consider to be sexist, others might not. This creates an additional layer of complexity for automated detection of sexism in text.

In AI-generated text, gender bias can result in toxic or stereotyped text outputs. Hence, multiple studies are dedicated to identifying this type of social bias in LLMs [6, 9, 43]. Most of the current methods propose automated evaluations based on a dataset to prompt an LLM and then measure if there are gender biases in the text outputs. More often than not, the existing approaches are gender-binary (e.g., men vs women), looking for imbalances in gender parity but ignoring diversity in the gender spectrum. Further, these techniques rely on ground truth metrics such as the Census breakdowns or annotated datasets, where an individual or group of individuals determine what is and what is not sexist [17]. However, as demonstrated by Mitamura et al. [34], purely objective agreement is not always possible when it comes to sexism, as some situations are open to interpretation, especially in nuanced, subtle sexism scenarios. In such cases, simply choosing just one interpretation as ground truth might result in a false sense that an LLM is unbiased when, in fact, it only responds to that single view and alienates other divergent opinions [16].

Consequently, in this paper, we argue for embracing subjectivity as a feature and not a flaw when it comes to analyzing gender biases in LLM outputs. Further, disagreements either between human evaluators or between human and automated evaluations, are to be expected in subtle and open-to-interpretation cases of sexism (and potentially other social biases as well). We discuss potential avenues to be explored in the next section.

3 FUTURE PATHS

The Natural Language Processing (NLP) community is the first place to draw inspiration from, since NLP researchers have extensively worked on automated hate speech detection [24]. Interestingly, the issue of subjectivity when evaluating subtle sexism (and social biases) in LLMs described in the previous section is comparable to the problem of low human annotator agreement identified in NLP research. The latter is found especially when labelling training datasets involving ambiguous tasks [1] such as identifying toxicity in language [46, 47] as well as in healthcare tasks [39] that do not have a ground truth. Gordon et al. [20] present an interesting solution to these *irreconcilable disagreements about ground truth* *labels*; instead of going for the majority vote, in detriment to minority voices, they propose balancing for which opinions to take into consideration, depending on context. This initiative could serve as inspiration to add subjective perspectives of human evaluators in addition to automated evaluation methods when identifying biases in LLMs. Further, while it can be more costly to include human evaluation as a methodology, [44]'s work could serve as inspiration in terms of best practices.

Another alternative for bias evaluation of LLMs stems from recent calls for bias management instead of aiming to completely debiasing an algorithm. Ferrara [15] highlights that as humans and human language exhibit social biases, it is expected that LLMs will reflect them as well, therefore wholly removing all biases might be an unfeasible task. In a bias management approach, as discussed in a recent article by Demartini et al. [8], biases are not entirely removed but instead surfaced to end-users, leveraging Explainable AI (XAI) techniques [32]. One potential avenue to be explored is leveraging human evaluation when end-users interact with a Large Language Model (LLM), similar to the one proposed by Shen et al. [42]. For instance, since sexism is inherently subjective and value-based (some end-users might consider an AI output as sexist, while others might not), embracing the subjective nature of gender bias should also be part of designing an LLM evaluation system. Hence, instead of aiming to calibrate the algorithm to an imperfect proxy of a ground truth (e.g., an annotated dataset with one single definition of what is or is not sexist) the system could elicit feedback from end-users in-situ, to then employ cultural adaptations as needed [25, 33, 37]. In practice, this could be formatted as user surveys while the end-user is interacting with the LLM to evaluate its outputs for biases. A design inspiration could be the existing spelling and grammar checking tools offered by Microsoft and Grammarly that also nudge users to use more inclusive language. This more human-centric evaluation approach also aligns with Zhu et al. [50]'s concept of value-sensitive algorithm design, advocating for pluralism in stakeholder needs that systems need to meet. There is further a potential for human-AI collaboration if the system works iteratively, incorporating user feedback on potential biases to improve and personalize outputs.

4 CONCLUSION

In this paper, we argue that automated techniques to evaluate if an LLM is biased represent an important advance towards more responsible AI applications. Nonetheless, as discussed in this work, there is opportunity to explore human evaluation and embrace subjectivity for a more holistic comprehension of the representation of social biases in these models and how they impact human-AI collaboration.

REFERENCES

[1] Sohail Akhtar, Valerio Basile, and Viviana Patti. 2020. Modeling annotator perspective and polarized opinions to improve hate speech detection. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 8. 151–154.

[2] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. 2021. Is the most accurate ai the best teammate? optimizing ai for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11405–11414.

[3] Manuela Barreto and Naomi Ellemers. 2005. The burden of benevolent sexism: How it contributes to the maintenance of gender inequalities. *European journal of social psychology* 35, 5 (2005), 633–642.

[4] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050* (2020).

[5] Laura Cabello, Anna Katrine Jørgensen, and Anders Søgaard. 2023. On the Independence of Association Bias and Empirical Fairness in Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 370–378.

[6] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[7] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109* (2023).

[8] Gianluca Demartini, Kevin Roitero, and Stefano Mizzaro. 2023. Data Bias Management. *arXiv preprint arXiv:2305.09686* (2023).

[9] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 862–872.

[10] Mark Diaz, Razvan Amironesei, Laura Weidinger, and Iason Gabriel. 2022. Accounting for Offensive Speech as a Practice of Resistance. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, Kanika Narang, Aida Mostafazadeh Davani, Lambert Mathias, Bertie Vidgen, and Zeerak Talat (Eds.). Association for Computational Linguistics, Seattle, Washington (Hybrid), 192–202. https://doi.org/10.18653/v1/2022.woah-1.18

[11] Kate Donahue, Alexandra Chouldechova, and Krishnaram Kenthapadi. 2022. Human-algorithm collaboration: Achieving complementarity and avoiding unfairness. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1639–1656.

[12] David Esiobu, Xiaoqing Tan, Saghar Hosseini, Megan Ung, Yuchen Zhang, Jude Fernandes, Jane Dwivedi-Yu, Eleonora Presani, Adina Williams, and Eric Michael Smith. 2023. ROBBIE: Robust Bias Evaluation of Large Generative Language Models. arXiv:2311.18140 [cs.CL]

[13] Percy Liang et al. 2023. Holistic Evaluation of Language Models. arXiv:2211.09110 [cs.CL]

[14] Sharon A Ferguson, Paula Akemi Aoyagui, and Anastasia Kuzminykh. 2023. Something Borrowed: Exploring the Influence of AI-Generated Explanation Text on the Composition of Human Explanations. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–7.

[15] Emilio Ferrara. 2023. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738* (2023).

[16] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2023. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770* (2023).

[17] Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling bias in toxic speech detection: A survey. *Comput. Surveys* 55, 13s (2023), 1–32.

[18] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462* (2020).

[19] Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic Bias Metrics Do Not Correlate with Application Bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 1926–1940. https://doi.org/10.18653/v1/2021.acl-long.150

[20] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

[21] Katherine J Hall. 1997. *Subtle Sexism: Current Practice and Prospect for Change.* Sage Publications.

[22] Katherine J Hall. 2016. *" They believe that because they are women, it should be easier for them." Subtle and Overt Sexism toward Women in STEM from Social Media Commentary.* Virginia Commonwealth University.

[23] Dong Huang, Qingwen Bu, Jie Zhang, Xiaofei Xie, Junjie Chen, and Heming Cui. 2023. Bias assessment and mitigation in llm-based code generation. *arXiv preprint arXiv:2309.14345* (2023).

[24] Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of Hate Speech automatic detection using Natural Language Processing. *Neurocomputing* (2023), 126232.

[25] Johanna Johansen, Tore Pedersen, and Christian Johansen. 2020. Studying the transfer of biases from programmers to programs. *arXiv preprint arXiv:2005.08231* (2020).

[26] Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2022. Debiasing isn't enough!–On the Effectiveness of Debiasing MLMs and their Social Biases in

Downstream Tasks. *arXiv preprint arXiv:2210.02938* (2022).

[27] Urja Khurana, Ivar Vermeulen, Eric Nalisnick, Marloes Van Noorloos, and Antske Fokkens. 2022. Hate speech criteria: A modular approach to task-specific hate speech definitions. *arXiv preprint arXiv:2206.15455* (2022).

[28] Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508* (2018).

[29] Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference*. 12–24.

[30] Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

[31] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards Understanding and Mitigating Social Biases in Language Models. arXiv:2106.13219 [cs.CL]

[32] Q Vera Liao and Kush R Varshney. 2021. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790* (2021).

[33] Kristian Lum, Yunfeng Zhang, and Amanda Bower. 2022. De-biasing "bias" measurement. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 379–389.

[34] Chelsea Mitamura, Lynnsey Erickson, and Patricia G Devine. 2017. Value-based standards guide sexism inferences for self and others. *Journal of Experimental Social Psychology* 72 (2017), 101–117.

[35] Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. *Sexuality & Culture* 25, 2 (2021), 700–732.

[36] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R Bowman. 2021. BBQ: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193* (2021).

[37] Katharina Reinecke and Abraham Bernstein. 2011. Improving performance, perceived usability, and aesthetics with culturally adaptive user interfaces. *ACM Transactions on Computer-Human Interaction (TOCHI)* 18, 2 (2011), 1–29.

[38] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301* (2018).

[39] Mike Schaekermann, Graeme Beaton, Elaheh Sanoubari, Andrew Lim, Kate Larson, and Edith Law. 2020. Ambiguity-aware AI Assistants for Medical Data Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (, Honolulu, HI, USA,) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376506

[40] Nikil Roashan Selvam, Sunipa Dev, Daniel Khashabi, Tushar Khot, and Kai-Wei Chang. 2022. The Tail Wagging the Dog: Dataset Construction Biases of Social Bias Benchmarks. *arXiv preprint arXiv:2210.10040* (2022).

[41] Nikhil Sharma, Q Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effects of LLM-Powered Search Systems on Diverse Information Seeking. *arXiv preprint arXiv:2402.05880* (2024).

[42] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–29.

[43] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326* (2019).

[44] Chris Van Der Lee, Albert Gatt, Emiel Van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*. 355–368.

[45] Lucía Vicente and Helena Matute. 2023. Humans inherit artificial intelligence biases. *Scientific Reports* 13, 1 (2023), 15737.

[46] Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*. 138–142.

[47] Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*. Association for Computational Linguistics, San Diego, California, 88–93. https://doi.org/10.18653/v1/N16-2013

[48] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876* (2018).

[49] Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't Make Your LLM an Evaluation Benchmark Cheater. *arXiv preprint arXiv:2311.01964* (2023).

[50] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on human-computer interaction* 2, CSCW (2018), 1–23.