Blind Spots and Biases: Exploring the Role of Annotator Cognitive Biases in NLP

Sanjana Gautam
Information Sciences and Technology
Pennsylvania State University, USA
sanjana.gautam@psu.edu

Mukund Srinath
Information Sciences and Technology
Pennsylvania State University, USA
mukund@psu.edu

Abstract

With the rapid proliferation of artificial intelligence, there is growing concern over its potential to exacerbate existing biases and societal disparities and introduce novel ones. This issue has prompted widespread attention from academia, policymakers, industry, and civil society. While evidence suggests that integrating human perspectives can mitigate bias-related issues in AI systems, it also introduces challenges associated with cognitive biases inherent in human decision-making. Our research focuses on reviewing existing methodologies and ongoing investigations aimed at understanding annotation attributes that contribute to bias.

1 Introduction

With the recent rapid expansion of generative AI models, we have witnessed their numerous benefits and the emergence of substantial ethical concerns (Thoppilan et al., 2022; Rudolph et al., 2023). There has been an influx of remarkable and noteworthy work that describes the issues of fairness, toxicity and bias in the text generation process (Bender et al., 2021; Abid et al., 2021; Seaborn and Kim, 2023). Several models are deployed as real-world solutions with a lack of informed consideration of their social implications, especially in sensitive fields such as healthcare, journalism, law, and finance (Khowaja et al., 2023). Recent research has revealed that these language models can mimic human biases present in language, perpetuating prejudiced behaviour that dehumanizes certain socio-demographic groups by deeming them more negative or toxic (Havens et al., 2022; Blodgett et al., 2020).

One of the proposed solutions to this issue has been to introduce human annotators to label the training corpora or validate pre-labelled datasets and manually remove toxic (or biased) data entries (Havens et al., 2022; Cabrera et al., 2014). It is common practice for machine learning systems to rely on crowd-sourced label data for training and evaluation (Wu et al., 2022). It is also well-known that biases present in the label data can induce biases in the trained models (Hettiachchi et al., 2021). Therefore, while humans-in-the-loop for model training may seem like an intuitive solution, it often introduces additional biases due to inherent cognitive biases in humans (Parmar et al., 2022). Crowdwork annotation studies conducted on MTurk (and other crowdwork platforms) where the participants come from a specific demographic population can potentially perpetuate populist viewpoints (Reinecke and Gajos, 2015).

Prior work has established the pitfalls in human rationality, as influenced by the lived experiences and environment, which Herbert Simon termed bounded rationality (Simon, 1957). Human biases have been identified to be the resulting gap between rational behaviour and heuristically determined behaviour (Tversky and Kahneman, 1974; Bojke et al., 2021). Over 180 cognitive biases have been identified, spawning everything from social interaction to judgment and decision-making with research spanning over 70 years (Talboy and Schneider, 2022). These tendencies or patterns can lead to faulty reasoning, irrationality, and potentially detrimental outcomes.

Bias sometimes emerges due to distractions, lack of interest, or laziness among annotators regarding the annotation task, leading them to select inaccurate labels. However, more concerning is the label bias stemming from informed and well-intentioned annotators who consistently exhibit disagreement (Hovy and Prabhumoye, 2021). Plank et al. (2014) demonstrated that this form of bias emerges when multiple correct labels are possible. For instance, the term 'social media' can be legitimately interpreted either as a noun phrase consisting of an adjective and a noun or as a noun compound comprising two nouns. For example, Sap et al. (2019) demonstrated that these biases mirror social and de-

Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing (3rd HCI+NLP Workshop), pages 82–88
June 21, 2024

mographic variances. For instance, annotators tend to evaluate utterances from various ethnic groups disparately and may misinterpret harmless banter as hate speech due to their unfamiliarity with the communication norms of the original speakers.

Merely relying on a few gold-standard corpora as training datasets or debiasing datasets is not a sustainable long-term strategy since languages undergo constant evolution. Thus, even a comprehensive sample can only encapsulate a momentary snapshot, offering at best a transient solution (Fromreide et al., 2014). We believe the design and set-up of the crowd work task plays a pivotal role in determining the goodness of data. In this work, we look at bias-diminishing strategies and identify the pressing questions in this area. Our central goal is to show that there is a need for standardized design principles when it comes to designing crowdwork studies. Specifically, we concentrate on the need for an HCI perspective in natural language processing research.

2 Bias in AI Models

Generative AI's propensity to amplify existing biases and create new ones has attracted considerable attention across a range of communities, including academics, policy-makers, industry, and civil society. Much of the initial work focused on developing quantitative definitions of fairness (Dwork et al., 2012; Hardt et al., 2016; Joseph et al., 2016; Liu et al., 2017; Verma and Rubin, 2018), and various technical methods for 'debiasing' AI models (Agarwal et al., 2018; Bolukbasi et al., 2016; Friedler et al., 2014; Zafar et al., 2017). When referring to de-biasing, we use the definition *'removing undesired skews in the data and the model outcome, such as by equalising a metric of interest between groups'*. *"Unintended bias"* is used to describe the different sources of bias that are introduced throughout an AI development life cycle (Lee and Floridi, 2021; Suresh and Guttag, 2021), focusing on not just the bias introduced, but also the harm it causes (Crawford, 2016).

Recent studies have shifted focus from merely identifying sources of bias in AI, such as flawed data collection methods, to exploring the various harms caused by these biases. This shift is supported by interdisciplinary research that highlights the contextual nature of fairness. Factors such as regional and cultural differences in lived experiences significantly influence perceptions of fairness, revealing that certain algorithmic behaviours may only be deemed harmful in specific social or cultural contexts (Green and Hu, 2018; Lee and Singh, 2021; Sambasivan et al., 2021; Selbst et al., 2019). Given these complexities, it is broadly acknowledged that eliminating bias or ensuring absolute fairness in AI systems is unfeasible (Kleinberg et al., 2016; Mehrabi et al., 2021; Pleiss et al., 2017). Instead, the objective is to minimize fairness-related harms and other adverse impacts to the greatest extent possible (Mehrabi et al., 2021; Selbst et al., 2019; Sun et al., 2019). This perspective is further enhanced by recent interdisciplinary studies (Lewicki et al., 2023), which underscore the nuanced and multifaceted nature of fairness in AI.

Identifying and acknowledging systemic biases in data collection is a crucial step in mitigating their impact on the systems that are trained using this data, and is a critical prerequisite for achieving fairness in algorithmic decision-making (Hajian et al., 2016). While humans are integral to the system, participating in data collection and various phases thereof, it is imperative to emphasize that human computation (Quinn and Bederson, 2011), the practice of harnessing human intelligence and cognitive abilities as computational elements, holds potential for addressing and mitigating these challenges.

3 Cognitive Bias among Annotators

As emphasized by Van Dis et al. (2023), ensuring human accountability is essential in scientific practice. The history of Large Language Models (LLMs) has shown that they can produce inaccurate information, or "hallucinations." To guarantee the accuracy of information, it is necessary to implement a rigorous verification and fact-checking process led by experts. Consequently, the discourse highlights the critical need for accountability in human-in-the-loop systems, particularly in response to the new challenges posed by these systems.

The importance of understanding and mitigating biases in crowd data is highly relevant to researchers, and others who rely on crowd data for creating automated systems. Prior work has explored various approaches to promoting fairness in machine learning, including the direct utilization of crowdsourced data (Balayn et al., 2018), leveraging crowds to assess perceived fairness of features (Van Berkel et al., 2019, 2021), applying pre-processing techniques such as removing sensitive attributes, resampling data to remove discrimination, and iteratively adjusting training weights for sensitive groups (Calmon et al., 2017; Kamiran and Calders, 2012; Krasanakis et al., 2018), as well as employing active learning methods (Anahideh et al., 2022).

The use of crowdsourcing for tasks such as data annotation can inadvertently introduce cognitive biases, stemming from the inherent design of the task itself. We have identified three primary reasons why annotated data can be problematic: (1) Unethical spammers submit imprecise or even arbitrary labels in order to maximize their financial advantage (Eickhoff et al., 2012) or due to external distractions. (2) Unqualified workers are, despite their best efforts, unable to produce an acceptable annotation quality (Eickhoff, 2014). (3) Malicious workers purposefully aim to undermine or influence the labelling effort (Wang et al., 2013). However, we propose that there might be some factors that have not been uncovered in prior literature. Crowd-workers have their tasks cut out for them, in cases where the nature of task design causes the propagation of bias. Research on crowd work has often focused on task accuracy whereas other factors such as biases in data have received limited attention (Hettiachchi et al., 2021).

Cognitive biases originate from individuals' own *"subjective social reality"* which is often a product of lived experiences. This makes cognitive bias a deviation from the rationality of judgement, therefore it may consist of perceptions of other people that are often illogical (Martie et al., 2005). An individual's construction of social reality, instead of the objective input, may dictate their behaviour and lead to perceptual distortion, inaccurate judgment, illogical interpretation, or irrationality (Bless and Fiedler, 2014). Past work has demonstrated that cognitive bias can affect crowdsourced labour and lead to significantly reduced result quality. This performance detriment is subsequently propagated into system ranking robustness and machine-learned ranker efficacy (Eickhoff, 2018).

The annotation instructions provided to crowdworkers can inadvertently prime them to exhibit biases towards or against specific domain information, which can be exacerbated by poorly designed instructions. Furthermore, annotators are often not fully informed about the true purpose of the research, leading to an ambiguity effect that can make the decision-making process appear more challenging and less appealing due to the limited information available (Ellsberg, 1961). Additionally, the phased revelation of information to annotators can result in an anchoring effect, where certain pieces of information are given disproportionate attention based on the timing of their disclosure. This underscores the importance of designing annotation studies that mitigate cognitive biases among workers, ensuring that the annotation process is fair, transparent, and unbiased.

4 Crowd Control

Humans in the loop bring a lot of value to generative AI and AI systems. Therefore, the solution to the issue of cognitive bias cannot be to remove the annotators from the system. Human annotators often bring expert judgements, that are valuable in creating ground truth labels. For example, annotation of medical imagery cannot be performed without the help of annotators who are medical professionals. Expert guidance, lived experiences and proximity to the problem domain make human annotators irreplaceable in the AI-training life-cycle. The common strategies of accounting for biases of annotators by employing qualification tests, demographic filters, incentives, and sophisticated worker models may not be enough to overcome this source of noise. There is therefore a need to control the annotation task design settings, to minimize the introduction of biases due to the cognitive biases of annotators. While cognitive biases and their effects on decision-making are well-known and widely studied, we note that AI-assisted decision-making presents a new decision-making paradigm. It is important to study their role in this new paradigm, both analytically and empirically.

Crowdwork platforms are often designed to position crowdworkers as interchangeable (Irani and Silberman, 2013). While some forms of digital work can be decomposed and distributed, the presumption that all crowdsourced dataset annotators exercise near-identical capacities of perception and judgement ignores the fact that social position, identity, and experience shape how annotators' actions.

Previous research has highlighted the significance of the annotator population and the power dynamics inherent in platform-mediated crowdwork, both of which can perpetuate cognitive biases (Díaz et al., 2022). Building upon this foundation, we propose a novel framework to enhance transparency and robustness in the process of designing a crowdwork task. This approach holds promise for mitigating the impact of cognitive biases in crowdwork, thereby contributing to more reliable and trustworthy outcomes.

5 Counter-measures for Biases

To minimize bias in NLP annotation tasks, several steps can be implemented. Firstly, recruiting a diverse group of annotators from various backgrounds can help balance individual biases. Providing clear and detailed guidelines ensures uniform understanding across annotators. Training sessions, followed by calibration discussions, align annotator interpretations and reveal guideline ambiguities. An iterative feedback loop allows for regular quality checks and guideline adjustments based on annotator experiences. Measuring inter-annotator agreement with metrics like Cohen's Kappa highlights discrepancies and areas needing clarification. Annotation tasks should be designed to minimize bias, such as by rotating text assignments among annotators to avoid topical biases. Finally, a post-annotation analysis can detect any remaining biases, ensuring the reliability and fairness of the annotated data.

However, biases can arise at any point in the AI lifecycle. It is therefore imperative for researchers to maintain a meticulous approach throughout the entire research process, encompassing various facets such as the selection of appropriate datasets, adherence to annotation schemes or labelling procedures, thoughtful considerations regarding data representation methodologies, judicious selection of algorithms tailored to the task at hand, and rigorous evaluation protocols for automated systems. Furthermore, researchers must consider the tangible real-world applications of their research endeavors. Particularly noteworthy is the imperative to consciously direct efforts towards leveraging technological advancements to uplift and empower marginalized communities, as underscored by Asad et al. (2019). Several studies critique existing bias mitigation algorithms for their lack of effectiveness due to inconsistent study protocols, inappropriate datasets, and over-tuning to specific test sets. To overcome these limitations, research needs to introduce robust evaluation protocol, and sensible metrics designed to evaluate algorithm robustness against various biases (Shrestha et al., 2022).

Our future work derives from the insights presented in the preceding discussion. It posits that the roots of bias within AI systems often traced back to the initial stages of the annotation process, particularly during the instruction phase. Although not all cognitive biases are inherently detrimental, a pressing need exists to advance our comprehension of how to devise annotation studies that align with the principles of human-computer interaction (HCI).

Our objective in this research endeavour is to contribute substantively to the ongoing efforts aimed at mitigating bias in crowd work. We intend to achieve this by focusing on the refinement of study design and instructional strategies. By incorporating insights from the HCI discipline, we aim to cultivate a nuanced understanding of how to create balanced annotation studies that minimize the emergence of bias. Through this work, we aspire to not only shed light on the pivotal role played by the annotation phase in propagating or mitigating bias but also to provide practical recommendations and guidelines for researchers and practitioners engaged in AI development and crowd work.

6 Conclusion

Our research highlights the critical importance of considering annotation attributes that contribute to bias in AI systems. The cognitive biases of annotators, inherent in human decision-making, can perpetuate and even amplify existing social disparities in AI models. To mitigate these issues, a multidisciplinary approach is necessary not only in deploying AI models but also in designing better systems for annotation tasks. By bringing together experts from diverse fields, including human-centered design, ethics, social sciences, law, healthcare, AI/ML, education, communication, and community representation, we can design annotation systems that are more inclusive, transparent, and fair. This collaborative framework is essential for developing annotation tasks that are free from biases, ambiguous, and unclear instructions, and that take into account the complexities of real-world data. Furthermore, a multidisciplinary approach is crucial for deploying AI models that are developed using these annotated data, ensuring that they are fair, transparent, and accountable. By acknowledging the limitations of human annotators and addressing them through a multidisciplinary approach, we can work towards a more equitable digital landscape where AI systems benefit both individuals and society as a whole.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *International conference on machine learning*, pages 60–69. PMLR.

Hadis Anahideh, Abolfazl Asudeh, and Saravanan Thirumuruganathan. 2022. Fair active learning. *Expert Systems with Applications*, 199:116981.

Mariam Asad, Lynn Dombrowski, Sasha Costanza-Chock, Sheena Erete, and Christina Harrington. 2019. Academic accomplices: Practical strategies for research justice. In *Companion Publication of the 2019 on Designing Interactive Systems Conference 2019 Companion*, pages 353–356.

Agathe Balayn, Panagiotis Mavridis, Alessandro Bozzon, Benjamin Timmermans, and Zoltán Szlávik. 2018. Characterising and mitigating aggregationbias in crowdsourced toxicity annotations. In *Proceedings of the 1st Workshop on Disentangling the Relation between Crowdsourcing and Bias Management*. CEUR.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Herbert Bless and Klaus Fiedler. 2014. *Social cognition: How individuals construct social reality*. Psychology Press.

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*.

Laura Bojke, Marta Soares, Karl Claxton, Abigail Colson, Aimée Fox, Christopher Jackson, Dina Jankovic, Alec Morton, Linda Sharples, and Andrea Taylor. 2021. Developing a reference protocol for structured expert elicitation in health-care decision-making: a mixed-methods study. *Health Technology Assessment (Winchester, England)*, 25(37):1.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Guillermo F Cabrera, Christopher J Miller, and Jeff Schneider. 2014. Systematic labeling bias: Debiasing where everyone is wrong. In *2014 22nd International Conference on Pattern Recognition*, pages 4417–4422. IEEE.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30.

Kate Crawford. 2016. Artificial intelligence's white guy problem. *The New York Times*, 25(06):5.

Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. 2022. Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2342–2351.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Carsten Eickhoff. 2014. Crowd-powered experts: Helping surgeons interpret breast cancer images. In *Proceedings of the First International Workshop on Gamification for Information Retrieval*, pages 53–56.

Carsten Eickhoff. 2018. Cognitive biases in crowdsourcing. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 162–170.

Carsten Eickhoff, Christopher G Harris, Arjen P de Vries, and Padmini Srinivasan. 2012. Quality through flow and immersion: gamifying crowdsourced relevance assessments. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 871–880.

Daniel Ellsberg. 1961. Risk, ambiguity, and the savage axioms. *The quarterly journal of economics*, 75(4):643–669.

Sorelle Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2014. Certifying and removing disparate impact. *arXiv preprint arXiv:1412.3756*.

Hege Fromreide, Dirk Hovy, and Anders Søgaard. 2014. Crowdsourcing and annotating ner for twitter# drift. In *LREC*, pages 2544–2547.

Ben Green and Lily Hu. 2018. The myth in the methodology: Towards a recontextualization of fairness in machine learning. In *Proceedings of the machine learning: the debates workshop*.

Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2125–2126.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Lucy Havens, Melissa Terras, Benjamin Bach, and Beatrice Alex. 2022. Uncertainty and inclusivity in gender bias annotation: An annotation taxonomy and annotated datasets of british english text. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 30–57.

Danula Hettiachchi, Mark Sanderson, Jorge Goncalves, Simo Hosio, Gabriella Kazai, Matthew Lease, Mike Schaekermann, and Emine Yilmaz. 2021. Investigating and mitigating biases in crowdsourced data. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, pages 331–334.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and linguistics compass*, 15(8):e12432.

Lilly C Irani and M Six Silberman. 2013. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 611–620.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems*, 29.

Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33.

Sunder Ali Khowaja, Parus Khuwaja, and Kapal Dev. 2023. Chatgpt needs spade (sustainability, privacy, digital divide, and ethics) evaluation: A review. *arXiv preprint arXiv:2305.03123*.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 world wide web conference*, pages 853–862.

Michelle Seng Ah Lee and Luciano Floridi. 2021. Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs. *Minds and Machines*, 31(1):165–191.

Michelle Seng Ah Lee and Jat Singh. 2021. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–13.

Kornel Lewicki, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. 2023. Out of context: Investigating the bias and fairness concerns of “artificial intelligence as a service”. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C Parkes. 2017. Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875*.

Haselton G Martie, Daniel Nettle, and Damian R Murray. 2005. The evolution of cognitive bias. 724-746 in the handbook of evolutionary psychology, edited by david m. buss.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2022. Don't blame the annotator: Bias already starts in the annotation instructions. *arXiv preprint arXiv:2205.00415*.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. *Advances in neural information processing systems*, 30.

Alexander J Quinn and Benjamin B Bederson. 2011. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1403–1412.

Katharina Reinecke and Krzysztof Z Gajos. 2015. Labinthewild: Conducting large-scale online experiments with uncompensated samples. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 1364–1378.

Jürgen Rudolph, Samson Tan, and Shannon Tan. 2023. Chatgpt: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, 6(1).

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 315–328.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th* *annual meeting of the association for computational linguistics*, pages 1668–1678.

Katie Seaborn and Yeongdae Kim. 2023. “i’m” lost in translation: Pronoun missteps in crowdsourced data sets. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–6.

Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 59–68.

Robik Shrestha, Kushal Kafle, and Christopher Kanan. 2022. An investigation of critical issues in bias mitigation techniques. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1943–1954.

Herbert A Simon. 1957. Models of man; social and rational.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Harini Suresh and John Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and access in algorithms, mechanisms, and optimization*, pages 1–9.

Alaina Talboy and Sandra Schneider. 2022. Reference dependence in bayesian reasoning: Value selection bias, congruence effects, and response prompt sensitivity. *Frontiers in Psychology*, 13:729285.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

Niels Van Berkel, Jorge Goncalves, Danula Hettiachchi, Senuri Wijenayake, Ryan M Kelly, and Vassilis Kostakos. 2019. Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–21.

Niels Van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B Skov. 2021. Effect of information presentation on fairness perceptions of machine learning predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Eva AM Van Dis, Johan Bollen, Willem Zuidema, Robert van Rooij, and Claudi L Bockting. 2023. Chatgpt: five priorities for research. *Nature*, 614(7947):224–226.

Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pages 1–7.

Tianyi Wang, Gang Wang, Xing Li, Haitao Zheng, and Ben Y Zhao. 2013. Characterizing and detecting malicious crowdsourcing. In *Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM*, pages 537–538.

Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180.