Biases and Ethical considerations for Machine Learning pipelines in the Computational Social Sciences

Suparna De, Shalini Jangra, Vibhor Agarwal, Jon Johnson and Nishanth Sastry

Abstract Computational analyses driven by Artificial Intelligence (AI)/Machine Learning (ML) methods to generate patterns and inferences from big datasets in computational social science (CSS) studies can suffer from biases during the data construction, collection and analysis phases as well as encounter challenges of generalizability and ethics. Given the interdisciplinary nature of CSS, many factors such as the need for a comprehensive understanding of different facets such as the policy and rights landscape, the fast evolving AI/ML paradigms and dataset specific pitfalls influence the possibility of biases being introduced. This chapter identifies challenges faced by researchers in the CSS field and presents a taxonomy of biases that may arise in AI/ML approaches. The taxonomy mirrors the various stages of common AI/ML pipelines: dataset construction and collection, data analysis and evaluation. With detecting and mitigating bias in AI an active area of research, this chapter seeks to highlight practices for incorporating responsible research and innovation into CSS practices.

Suparna De
University of Surrey, UK. e-mail: s.de@surrey.ac.uk

Shalini Jangra
University of Surrey, UK. e-mail: s.jangra@surrey.ac.uk

Vibhor Agarwal
University of Surrey, UK. e-mail: v.agarwal@surrey.ac.uk

Jon Johnson
University College, London (UCL), UK. e-mail: jon.johnson@ucl.ac.uk

Nishanth Sastry
University of Surrey, UK. e-mail: n.sastry@surrey.ac.uk

1 Introduction

Advances in communication networks and the growing use of social networking platforms means that there is an unprecedented amount of information that provides an important source for understanding a population [1, 2]. Computational tools have been successfully used to analyze the resulting structured and unstructured data, with the aim of understanding individuals, groups and their social practices. This well-studied field of computational social science (CSS) is characterized by: (1) the involvement of human subjects, with the resulting capabilities and tools also impacting individuals and communities, (2) the use of large and complex datasets, drawn from mixed methods data collection, incorporating both self-reporting through surveys and experiments, as well as through observation of 'unconstrained' behaviour on social media platforms, (3) application of AI or ML-driven computational or algorithmic solutions to the resulting big data to generate insights, inferences and predictions about human behaviours, social networks and systems.

We cannot use ML predictive models in a black box fashion for social science problems [24]. It is necessary to analyze the ethical implications and consequences of these models' output as these may have real world consequences and impacts. Due to this human impact, computational research needs to be "ethical, trustworthy and responsible" [3]. However, this very human nature of the data means that it encounters issues of representativeness, uniformity and bias [1]. Thus, this chapter focuses on some of the key issues around ethics and generalizability confronting CSS researchers in the age of big data. These issues are analyzed through the lens of the data lifecycle in ML pipelines, as identified in existing literature [4], i.e. covering dataset creation/collection, data analysis and data (model) evaluation, as shown in Figure 1. This is followed by a discussion of the strategies and existing initiatives to address the issue of bias in CSS ML pipelines.

2 Dataset Creation and Collection Bias

The first stage of a typical ML pipeline starts with data collection, which can take the form of scraping it from social networking platforms, e.g., Reddit [39, 38] and Kialo [36, 37] or creating a dataset from available survey data collection APIs [5]. The creation and archiving of such complex datasets naturally gives rise to issues of data privacy and de-identification, necessitating steps for individual privacy protection and conforming to laws and principles of informed consent (e.g. GDPR[1]).

The following sub-sections describe how biases can be introduced in the ML pipeline during the dataset creation and collection phase, which also includes labelling or annotating the data.

[1] https://eugdpr.org/

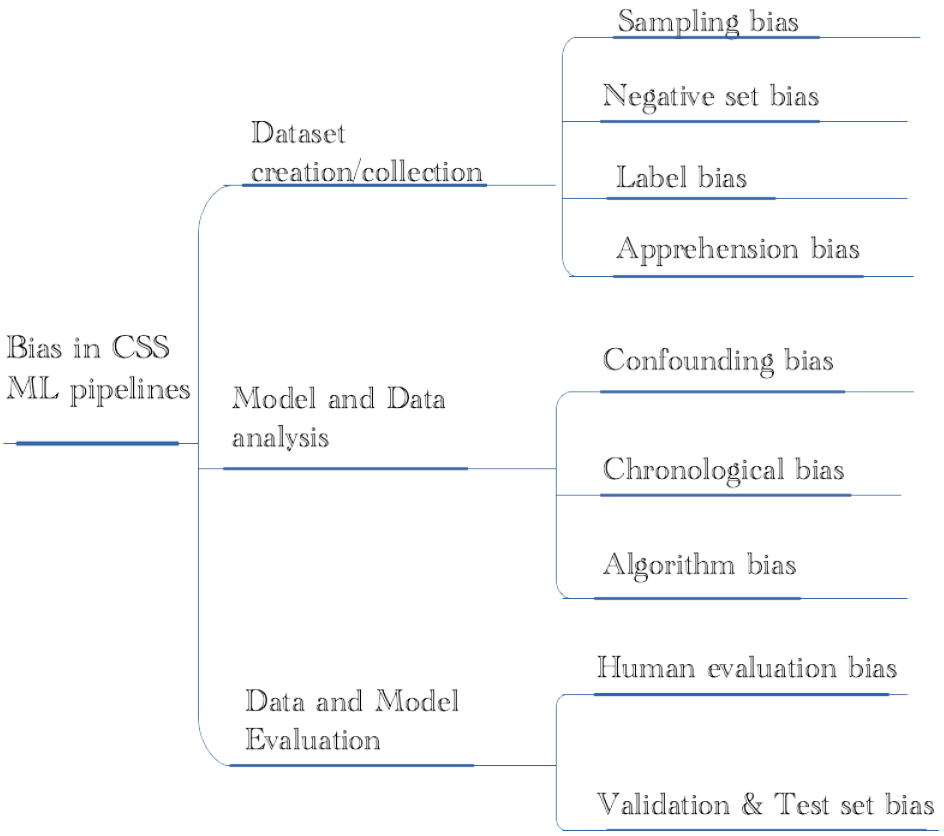

Fig. 1 Taxonomy of biases in CSS ML pipelines

2.1 Sampling Bias

One of the most common instance of dataset bias is sampling bias, which occurs due to some types of instances being selected more than others [4]. Datasets are often created with a particular set of instances, with most social media research using a sample of all available data to make inferences about a larger population [47]. With the sampling methods necessitating representativeness of both demographics and behaviour, any systemic distortion in the sampled data, due to sparsity for instance, can compromise its representativeness. It is also difficult to obtain a uniform random sampling from social platforms. Sparsity in the data can also be magnified due to platform characteristics, for instance, by limiting the length of users' posts which in turn affect data retrieval [50]. Therefore, poor generalization of the trained AI models can be an unintended consequence of sampling bias.

The probability of gaps in data coverage also increase in the case of longitudinal studies spanning decades, as in the case of the UK's MRC National Survey of Health

and Development (MRC/NSHD) study, which has a lower occurrence of labelled instances in some vocabulary categories such as measures of psychological well-being, omics and sleep, compared to more recent birth cohort studies [6].

This sampling bias can also occur due to missing instances or features in the datasets and socio-cultural conditions of data generation. The importance of context around the social and historical conditions in the data generation process is also crucial where observational data may have 'non-random missingness' [3] and meaningful noise.

2.2 Negative Set Bias

Negative set bias occurs when there are not enough samples representative of the remaining world (negative instances which are not present in the dataset). The ever-increasing use of social platforms and behaviour capture by both private corporations and government bodies has led to unprecedented amounts of data being collected on human activities' traces. However, historically disadvantaged groups are often "less involved in the formal economy and its data generating activities" [8], which means that there are not enough samples representing such groups in the dataset, causing negative set bias. This leads to potential reinforcement of digital divides and data inequities through biased techniques that render digitally marginalised groups invisible.

Negative set bias may also be manifested due to user *self-selection* bias, either due to users exercising self-censorship [51, 53], e.g. not 'liking' or sharing/deleting a post despite reading it, due to privacy concerns. It can also occur due to platform characteristics which makes some user activities invisible, e.g. dataset only includes users who post content, not those who only read it [52].

A related ethical question is that most of the data harvesting occurs without the conscious "consent or active awareness of the people whose digital and digitalised lives are the targets of surveillance, consumer curation, and behavioural steering" [3], raising questions of privacy, autonomy and meaningful consent. An example can be found in the geo-tagging capabilities of some social networking platforms, with some users unaware of their posts being geotagged, while others consciously using geotagging to 'advertise' where they have been [54]. Negative set bias also has real-world implications when the resultant analyses are used to inform data-driven public policies, which may be geared towards economically-advantaged and data-rich areas [55].

The opposite of this 'negative' bias is in domains such as hate speech, where there are insufficient positive samples (most datasets have very few hate speech occurrences). This was a problem for us in the work with MPs [70] and also more recently in the Decentralised Web [71]. Vidgen *et al.* [72] have a unique approach to this problem - they artificially generate (through crowd workers) a balanced data set on hate speech.

2.3 Label Bias

Label bias is bias associated with the labelling or data annotation process. Subjective biases and domain background of the annotators can deeply influence the annotation process, leading to inconsistencies in the labelling process. Different annotators have different perspectives based on their different life experiences and world view [40]. For example, annotating hate speech is a highly subjective task [41]. Often, different annotators give different labels to instances based on their varying levels of sensitivity towards a particular hate type. Their aggregated labels, mostly using majority voting, are often treated as gold labels in various hate speech datasets and therefore, favour majority opinions [42]. ML models trained over these datasets with label biases can be highly biased in nature and can result in poor performance in detecting hate speech accurately. Guest *et al.* [69] replace majority voting with facilitated meetings between annotators to improve the quality of the datasets generated. AnnoBERT [73] directly incorporates subjectivity into a hate speech detection model and shows that this improves classification performance.

The subjectivity of the labelling process also contributes to its propensity towards bias, which can be magnified in the case of high-volume longitudinal studies, as reported in our recent work [6], as the labels given for an object type can diverge significantly more than where the data collection period is short. As reported in this work, unsupervised topic modelling approaches uncovered instances of unintuitive manual labelling in cases of semantic overlaps in question texts, with the mislabelled instances reflecting the domain background of the human labellers.

2.4 Apprehension Bias

Apprehension bias is concerned with how user behaviour (and hence, how it is manifested in the resulting dataset) is impacted by the awareness of being observed. In response to observers such as other platform users or administrators, users may choose different behaviours of *self-presentation*, which is termed as online "Hawthorne effect" [47]. Such effects have been studied in location-based social networks, where check-ins at public locations such as restaurants are more likely than at private ones such as a doctor's surgery [61]. Conditioning of individual writing style of reviews has been found to be influenced by prior ratings and reviews [62].

Apprehension bias is also prominent in observational CSS, where study participants are recruited for administering surveys or questionnaires, as this brings into play the researchers as active observers, which can cause a behavioural change as a conscious response to being studied. This is illustrated in the mixed-mode data collection stage for the National Child Development Study survey, as reported in [7]. The authors of this work report not only variance in participation rates between telephone-only and Web-based respondents to the survey, but also differences in response values which can be attributable to the mode of data collection. For instance, Web-based participants had a higher non-response to questions related to finances,

and also had more negative stances to self-rated subjective parts of the study, such as health and well-being. As a result, the authors identified the potential for subsequent biases in the analyses, and recommended techniques to correct for these.

3 ML Model and Data Analysis Bias

A second realm of problems concerns the construction of the algorithms (if they are structured and not completely self-learning), and the selection of features or criteria. Biases can be introduced through untrue assumptions of the distribution of the data, the data cleaning and pre-processing methods as well as the choice of the ML models.

3.1 Confounding Bias

Confounders are external variables that manipulate the estimate of the apparent relationship between the independent variable of interest and the dependent (output) variable and hence lead to erroneous output of the model [26]. A confounding variable can influence the outcome of an experiment in various ways, such as: invalid correlations, increasing variance and suggesting an association where none exists or masking a true association. Confounding, sometimes referred to as confounding bias, is mostly described as a "mixing or blurring of effects" [27]. For instance, [32] states that the root reason for the bias in recommender systems present in e-commerce (e.g. Amazon and Alibaba) websites and social networking platforms such as Twitter or Facebook are confounder variables that influence both which items the user will interact with and how they rate them. Approaches to address the detrimental effects of confounding variables include those by Liu et al. [33] who proposed a debiased information bottleneck (DIB) objective function to reduce the confounding bias in the biased feedback without having to retrain with unbiased data. Randomization such as random initialization or random choices during learning is the only way to control for confounding because it will balance measured and unmeasured confounding.

A type of confounding bias is that of 'omitted variable', where the analysis is carried out without considering the relevant features. This is more significant for predictive ML, such as regression analysis, when the omitted variables match the independent variables or regressors and the dependent variables are determined by this omitted variable [43]. This causes the analysis to correlate their effects to model variables that caused bias, to the estimated effects, thus, confounding the cause-effect relationship, making it challenging to differentiate between "attributes that merely correlate and those that are causally related" [47]. An example is the spurious correlation between URLs in tweets and their retweet rates, which were found to be due to the URLs often co-occuring with hashtags [65]. Consequences of omitted variable bias include both exaggerating and underrating the effect in the

analysis, flipping the statistical analysis result or even causing an effect to be hidden in the outcome.

A related concept to omitted variable bias is 'proxy' or indirect bias, with variables used as proxies for sensitive ones, or those that are not directly measurable. The use of proxy variables abounds in CSS analyses, though they may suffer from validity or reliability issues [47]. For social networks research, interest in a topic is often indirectly measured through the proxy variable of number of posts on the topic [46], though it fails to conclusively capture how much content of the topic is actually read. The choice of proxy participants to determine user traits or demographic criteria has also been shown to influence the performance of prediction models, for example, in the case of using university alumni registered on a social platform as proxy for 'young' college graduates to determine their views on a new law [48], which resulted in an important source of bias.

3.2 Chronological Bias

Chronological bias refers to the change in study design that happens over time and effects the study results, due to temporal variations caused by population drifts or system drifts [56].

System drifts can lead to issues of 'temporal validity' of the study conclusions as illustrated in the case of the Google Flu Trends (GFT) platform, which following an algorithm update in 2009, made headlines in 2013 for predicting more than double the number of doctor visits for flu-like illness versus that reported by the Centers for Disease Control and Prevention (CDC). An analysis into the GFT over-estimation [49] revealed issues with the algorithm dynamics and changes made in the underlying Google search algorithm in June 2011 and February 2012. The analysis uncovered that Google's modifications in search results in 2011/12, to suggest additional search terms and also potential diagnoses for searches, tracked closely with GFT errors when comparing correlated search terms for the GFT time series to those returned by the CDC data.

Population drifts occur when study participants whose data is mined or analysed earlier during an intervention are subject to different social exposures or are at a different risk from participants who are recruited later [9]. This has been exemplified with studies on both the Facebook [57] and Twitter [58] social platforms. Changes in platform users' lifestyles and evolution of online communities [59] can also affect how long users are engaged with a topic, which may also be dependent on changes in the platform itself, such as the addition of new features.

3.3 Algorithm Bias

Algorithm bias is defined as bias that is solely induced or added by the algorithm, for instance, a ML model that relies on randomness for fair distributions of results is not truly random.

Specific types of such bias include *ranking* bias - privileging some algorithmic results more than others in the way they are presented. For instance, social media platforms employ algorithms designed to promote trending content that may negatively affect the overall quality of information on the platform. As an extension to ranking bias, personalisation algorithms employed in social media platforms and search engines are designed to select only the most engaging and relevant content for each individual user. But in doing so, it "may end up reinforcing the cognitive and social biases of users" [60], with less diverse exposure to content, thus making them part of a social bubble and more vulnerable to manipulation.

Another case is of *insensitive measure* bias [9] that can result from the use of an insufficiently accurate method to detect the outcome of interest, where the method is not sensitive enough to detect true differences. Examples include use of automated Natural Language Processing (NLP) tools for dependency parsing and language detection, which may not be robust when different dialects, which vary from the mainstream languages, are present in the dataset [64]. The use of alternative objective functions, when the true criterion is not directly measurable, such as user clicks as a substitute for user satisfaction [66], have the potential of creating 'Matthew effects' of self-reinforcing feedback loops [8] between datasets, decisions and algorithms. Such effects can have harmful downstream consequences such as false negatives disappearing from the dataset [8], with the resulting asymmetry skewing the decision-making process.

Social media platforms also expose users to a less diverse content from a significantly narrower spectrum of sources compared to non-social media sites like Wikipedia [44]. This is called as *homogeneity* bias. This can take the form of 'gatekeeping', where there is a distinct preference for some topics, and 'coverage', with differences in attention given to certain topics as well as how these are presented [63]. Pre-trained language representations such as Bidirectional Encoder Representations from Transformers (BERT), which is trained on a general-purpose corpus may under- or over-represent the relationship between different words in the dataset under analysis, as even though different scientific domains may use the same language, the words may have very different semantic connotations.

4 Data and Model Evaluation Bias

4.1 Human Evaluation Bias

Biases during evaluation can be introduced by circumstances of *confirmation bias* (interpreting information that is consistent with existing beliefs), *peak end effect* (cognitive bias related to how subjects remember, by focussing the recall on the 'peak' or more intense moments), and prior beliefs (e.g., culture). For instance, the detailed advertising tools built into many social networking platforms play into the hands of actors looking to spread disinformation by tailoring messages to people who are already inclined to believe them, i.e. exploiting confirmation bias [45]. Human evaluators are also limited by how accurately or by how much information they can recollect, which can result in *recall bias*. People show this bias when they reminiscence information selectively (by omitting details), or when they understand/assess it in a biased way. Schwind *et al.* [28] state that the pattern of results observed in selection behavior is also apparent in evaluation behavior. This implies that the reduction of evaluation bias will occur only when preference-inconsistent recommendations are combined with low prior knowledge conditions.

4.2 Validation and Test Set Bias

Validation and test set bias refer to systematically under- or over-estimating the predictive performance of the model [68]. Practitioners introduce bias into their model when tuning new models based on the performance of old models on the test/holdout data. For instance, developers make changes to the model based on what they have learned about how previous topologies and hyperparameters affected the model accuracy on the test data, thereby introducing bias into the model. By leveraging observations gained from the model's performance on test data, it is possible to optimize the model using the whole dataset, avoiding ever training the model straight on the test samples. The presence of bias in models can be influenced by the samples and labels chosen in the validation and test datasets [4]. Evaluation bias can also arise from inadequate benchmarks or datasets used for testing. Consequently, metrics computed over the whole test or validation set may not always provide a correct indication of the model's fairness.

5 Responsible Research for CSS ML Pipelines

Several initiatives exist, such as 'Datasheets for Datasets' [10] for documenting essential information about datasets for training ML models, as part of a move towards practical guidelines for reducing potential bias in AI systems. Others aimed

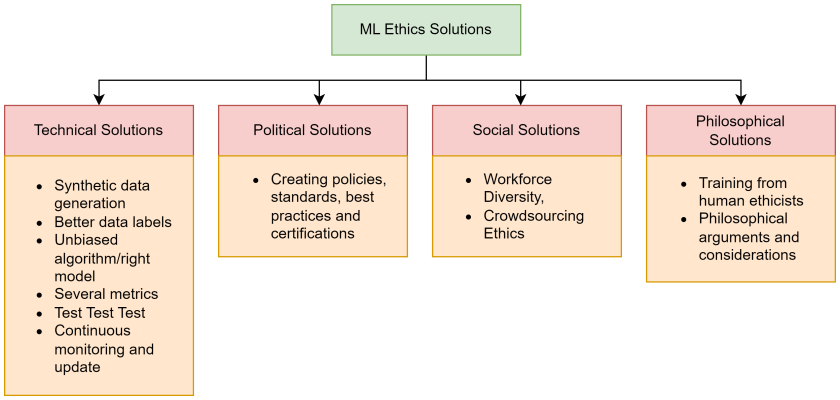

Fig. 2 Ethical Solutions to reduce Bias in Machine Learning

at a technical level include initiatives such as 'discrimination-aware data mining' (DADM) [8].

Strategies for choosing ML models that may be less discriminatory than baseline choices can include adversarial debiasing [11], where the model learns to predict the outcomes to prevent another adversary AI model from guessing the protected variables based on the outcomes. Another strategy is the dynamic upsampling of training data [12], with the data from underrepresented groups being given more weight during the training phase.

Approaches for reducing CSS bias can be divided into three categories: i) pre-processing approaches [14], ii) in-processing approaches [16, 15] and iii) post-processing approaches [17]. Pre-processing approaches target the foremost source of bias i.e. data. Their prime objective is to generate a balanced and fair dataset that results in less discriminative ML models [13]. These approaches include altering the data distribution by sampling, re-weighting, or modifying the individual training instance. Modeling classification problems with fairness constraints [18], restricting the learner's behavior by enforcing independence on sensitive features [19] and adversarial debiasing [11] are some of the different in-processing bias mitigation approaches. Post-processing approaches are applied once the model has been trained on the data, which includes changing the model's internals (white-box approaches) [22, 23] or its predictions (black-box approaches) [17, 20, 21]. Bias-mitigation approaches should provide the middle ground between the ML model's accuracy and fairness.

The solutions proposed for ethical approaches to reduce CSS biases are of four types: technical, social, political, and philosophical [29], as depicted in Figure 2.

1. **Technical Solutions**: One of the prime technical solutions for mitigating bias is synthetic data generation. Generating synthetic data involves defining and setting the parameters of a fair dataset and then generating data that fulfills that definition, which may help protect people's sensitive information. Mislabeling due to preconceived notions and assumptions of labelers can have unintentional or

detrimental real-world consequences. More nuanced labels or categories can help introduce fairness in the system. Additional contextual metadata, for example, the characteristics of the population and the mode of data collection, may also be used to identify and mitigate potentially unmeasured biases. Users' configuration of the algorithm could reflect their cultural and experiential biases. Therefore, having absolute transparency on how the algorithm works can be helpful in designing unbiased algorithms. Train then mask emphasizes helping marginalized groups while treating the non-sensitive features as the same as others [31]. Setting up the correct parameters, regular spot checks and continuous testing and monitoring also help.

2. **Political Solutions**: It is required to establish political control over the ethical deployment of AI/ML to reduce the likelihood of unintended consequences. This can be achieved by creating guidelines, policies & legal frameworks and introducing certifications to learn best practices for the responsible use of these technologies. For instance, the EU passed the General Data Protection Regulation (GDPR) in April 2016 which came into effect in 2018. It mandates organizations to provide citizens in the EU with the "right to explanation," which refers to the right to receive an explanation for an algorithm's output. The government's investment in ethical ML technologies research can help to build a knowledgeable workforce capable of developing and deploying ethical machine learning systems.
3. **Social Solutions**: Raising awareness among the public can be an approach to tackling ML bias. For instance, Google and Microsoft researchers founded the workshop "Fairness, Accountability, and Transparency in Machine Learning" (FAT ML) to examine the repercussions of algorithmic bias. It has now developed into the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)[2], which brings together researchers and practitioners from across computer science, law, social sciences, and humanities to tackle issues in this area. The involvement of people from diverse populations in the entire ML pipeline will also reduce discrimination. Radford *et al.* [67] argue that biases emerge throughout the entire ML pipeline that cannot be remedied solely through technical solutions. They describe the way social theory, for e.g., critical race theory and feminist theory, can help in removing ML bias by providing a framework for understanding the social and cultural contexts in which the data is produced and used.
4. **Philosophical Solutions**: Considering all contextual divergences, humans are the final piece to making ethical decisions. Although machine learning may produce practical and advanced applications, more is needed to replace the human capacity for domain expertise. Gnjatović *et al.* [34] focused on reintroducing humans into the learning loop. For instance, false, inaccurate, or incomplete information floated online in news, social media, and on the Web causes societal harm. Reference [35] discusses the importance of hybrid approaches to fighting against online misinformation and disinformation. Both ML tools and humans - including spe-

[2] https://facctconference.org

cialized professionals and lay persons sourced through crowdsourcing platforms, should collaborate to mitigate the issue.

Acknowledgements This research is funded by the UKRI Strategic Priority Fund as part of the wider Protecting Citizens Online programme (Grant number: EP/W032473/1) associated with the National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online (REPHRAIN), and by the Science and Technology Facilities Council (STFC) DiRAC-funded 'Understanding the multiple dimensions of prediction of concepts in social and biomedical science questionnaires' project, grant number ST/S003916/1.

References

1. Shah, D. V., Cappella, J. N., and Neuman, W. R.: Big Data, Digital Media, and Computational Social Science: Possibilities and Perils. The Annals of the American Academy of Political and Social Science. **659**(1) 6–13 (2015) doi: 10.1177/0002716215572084
2. De, S., Jassat, U., Grace, A., Wang, W., and Moessner, K.: Mining Composite Spatio-Temporal Lifestyle Patterns from Geotagged Social Data In: 2022 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics), Espoo, Finland (2022) pp. 444-451.
3. Leslie, D.: Don't "research fast and break things": On the ethics of Computational Social Science. (2022) ArXiv, abs/2206.06370.
4. Ramya Srinivasan, R. and Chander, A.: Biases in AI Systems: A survey for practitioners. ACM Queue **19** (2) (March-April 2021)
5. De, S., Moss, H., Johnson, J., Li, J., Pereira, H., & Jabbari, S.: Engineering a machine learning pipeline for automating metadata extraction from longitudinal survey questionnaires. IASSIST Quarterly, **46**(1) (2022)
6. Sharifian-Attar, De, S., Jabbari, S., Li, J., Moss, H., Johnson, J., : Analysing Longitudinal Social Science Questionnaires: Topic modelling with BERT-based Embeddings. In: Proc. 2022 IEEE International Conference on Big Data, Osaka, Japan, 2022, pp. 5558-5567, doi: 10.1109/BigData55660.2022.10020678.
7. Goodman, A., Brown, M., Silverwood, R. J., Sakshaug, J. W., Calderwood, L., Williams, J., Ploubidis, George B. : The impact of using the Web in a mixed-mode follow-up of a longitudinal birth cohort study: Evidence from the National Child Development Study. Journal of the Royal Statistical Society: Series A (Statistics in Society), **185**(3) 822–850 (2022)
8. Herzog, L.: Algorithmic Bias and Access to Opportunities. In: Véliz C. (ed.), The Oxford Handbook of Digital Ethics (2021) doi: 10.1093/oxfordhb/9780198857815.013.21
9. Spencer E.A., and Heneghan C.: Catalogue of Bias Collaboration. In: Catalogue Of Bias (2017). https://catalogofbias.org/biases/
10. Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé III, H., and Crawford, K.: Datasheets for datasets. Commun. ACM 64, 12 (2021), 86–92 doi: 10.1145/3458723
11. Zhang, B. H., Lemoine, B, and Mitchell, M.: Mitigating Unwanted Biases with Adversarial Learning. In: Artificial Intelligence, Ethics, and Society Conference (2018)
12. Cofone, I. N.: Algorithmic Discrimination Is an Information Problem.Hastings Law Journal **70** 1389–1444 (2019)
13. Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., ... & Staab, S.:Bias in data-driven artificial intelligence systems—An introductory survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, **10** (3), e1356 (2020)
14. Hajian, S.:Simultaneous discrimination prevention and privacy protection in data publishing and mining. arXiv preprint arXiv:1306.6805 (2013)

15. Fish, B., Kun, J., & Lelkes, Á. D.: A confidence-based approach for balancing fairness and accuracy. In Proceedings of the 2016 SIAM international conference on data mining (pp. 144-152). Society for Industrial and Applied Mathematics (2016, June)
16. Kamishima, T., Akaho, S., & Sakuma, J.: Fairness-aware learning through regularization approach. In 2011 IEEE 11th International Conference on Data Mining Workshops (pp. 643-650). IEEE (2011)
17. Hardt, M., Price, E., & Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems, **29** (2016)
18. Celis, L. E., Huang, L., Keswani, V., & Vishnoi, N. K. (2019, January). Classification with fairness constraints: A meta-algorithm with provable guarantees. In Proceedings of the conference on fairness, accountability, and transparency (pp. 319-328)
19. Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012, September). Fairness-aware classifier with prejudice remover regularizer. In Joint European conference on machine learning and knowledge discovery in databases (pp. 35-50). Springer, Berlin, Heidelberg
20. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018, July). A reductions approach to fair classification. In International Conference on Machine Learning (pp. 60-69). PMLR.
21. Canetti, Ran, Aloni Cohen, Nishanth Dikkala, Govind Ramnarayan, Sarah Scheffler, and Adam Smith. "From soft classifiers to hard decisions: How fair can we be?." In Proceedings of the conference on fairness, accountability, and transparency, pp. 309-318. 2019.
22. Pedreschi, D., Ruggieri, S., & Turini, F. (2009, April). Measuring discrimination in socially-sensitive decision records. In Proceedings of the 2009 SIAM international conference on data mining (pp. 581-592). Society for Industrial and Applied Mathematics.
23. Calders, T., Kamiran, F., & Pechenizkiy, M. (2009, December). Building classifiers with independency constraints. In 2009 IEEE International Conference on Data Mining Workshops (pp. 13-18).
24. Wallach, H. (2018). Computational social science ≠ computer science + social data. Communications of the ACM, **61** (3), (pp. 42-44).
25. Garcia, M. (2017). Racist in the Machine: The Disturbing Implications of Algorithmic Bias. World Policy J **33** (4), (pp. 111–117)
26. Zhao, Q., Adeli, E., & Pohl, K. M. (2020). Training confounder-free deep learning models for medical applications. Nature communications, **11** (1), (pp. 1-9).
27. Jager, K. J., Zoccali, C., Macleod, A., & Dekker, F. W. (2008). Confounding: what it is and how to deal with it. Kidney international, **73** (3), (pp. 256-260).
28. Schwind, C., & Buder, J. (2012). Reducing confirmation bias and evaluation bias: When are preference-inconsistent recommendations effective–and when not?. Computers in Human Behavior, **28** (6), (pp. 2280-2290).
29. Shadowen, N. (2019). Ethics and bias in machine learning: A technical study of what makes us "good". In The Transhumanism Handbook (pp. 247-261). Springer, Cham.
30. Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., & Sculley, D. (2017). No classification without representation: Assessing geodiversity issues in open data sets for the developing world. arXiv preprint arXiv:1711.08536.
31. Ghili, S., Kazemi, E., & Karbasi, A. (2019, July). Eliminating latent discrimination: Train then mask. In Proceedings of the AAAI Conference on Artificial Intelligence **33**, (01), pp. 3672-3680).
32. He, M., Hu, X., Li, C., Chen, X., & Wang, J. (2022). Mitigating Confounding Bias for Recommendation via Counterfactual Inference. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD22).
33. Liu, D., Cheng, P., Zhu, H., Dong, Z., He, X., Pan, W., & Ming, Z. (2021, September). Mitigating confounding bias in recommendation via information bottleneck. In Fifteenth ACM Conference on Recommender Systems (pp. 351-360).
34. Gnjatović, M., Maček, N., & Adamović, S. (2020). Putting Humans Back in the Loop: A Study in Human-Machine Cooperative Learning. Acta Polytechnica Hungarica, 17(2).

35. Demartini, G., Mizzaro, S., & Spina, D. (2020). Human-in-the-loop Artificial Intelligence for Fighting Online Misinformation: Challenges and Opportunities. IEEE Data Eng. Bull., 43(3), 65-74.
36. Agarwal, V., Joglekar, S., Young, A. P., & Sastry, N. (2022). GraphNLI: A Graph-based Natural Language Inference Model for Polarity Prediction in Online Debates. In Proceedings of the ACM Web Conference 2022 (pp. 2729–2737).
37. Young, A. P., Joglekar, S., Agarwal, V., & Sastry, N. (2022). Modelling online debates with argumentation theory. ACM SIGWEB Newsletter, (Spring), (pp. 1–9).
38. Agarwal, V., Young, A. P., Joglekar, S., & Sastry, N. (2022). A Graph-Based Context-Aware Model to Understand Online Conversations. arXiv preprint arxiv.2211.09207.
39. Guest, E., Vidgen, B., Mittos, A., Sastry, N., Tyson, G., & Margetts, H. (2021). An expert annotated dataset for the detection of online misogyny. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (pp. 1336–1350).
40. Akhtar, S., Basile, V., & Patti, V. (2020). Modeling annotator perspective and polarized opinions to improve hate speech detection. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (pp. 151–154).
41. Aroyo, L., Dixon, L., Thain, N., Redfield, O., & Rosen, R. (2019). Crowdsourcing subjective tasks: the case study of understanding toxicity in online discussions. In Companion proceedings of the 2019 World Wide Web conference (pp. 1100–1105).
42. Sheng, V. S., Zhang, J., Gu, B., & Wu, X. (2017). Majority voting and pairing with multiple noisy labeling. IEEE Transactions on Knowledge and Data Engineering (pp. 1355–1368).
43. Wilms , R., Mäthner, E., Winnen, L., Lanwehr, R.. (2021). Omitted variable bias: A threat to estimating causal relationships. Methods in Psychology **5** 2021.
44. Nikolov, D., Oliveira, D. F., Flammini, A., & Menczer, F. (2015). Measuring online social bubbles. PeerJ computer science, 1, e38.
45. Ciampaglia, G. L., & Menczer, F. (2018). Misinformation and biases infect social media, both intentionally and accidentally. The Conversation, 20.
46. Chen, J., Nairn, R., Nelson, L., Bernstein, M., and Chi, E. Short and tweet: experiments on recommending content from information streams. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10), New York, NY, USA, 1185–1194 (2010).
47. Olteanu A., Castillo, C., Diaz, F., & Kıcıman, E. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. Frontiers in Big Data, **2** (2019).
48. Cohen, R., & Ruths, D. Classifying Political Orientation on Twitter: It's Not Easy!. In: Proceedings of the International AAAI Conference on Web and Social Media, **7** (1), 91-99. (2013).
49. Lazer D., Kennedy, R., King, G., Vespignani, A. The Parable of Google Flu: Traps in Big Data Analysis. Science, **343** (6176), 1203-1205. (2014).
50. Naveed, N., Gottron, T., Kunegis, J., and Alhadi, A. C. Searching microblogs: coping with sparsity and document quality. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM '11, New York, 183–188. (2011).
51. Gong,W., Lim, E.-P., Zhu, F., and Cher, P. H. On unravelling opinions of issue specific-silent users in social media. In: Proceedings of the International AAAI Conference on Web and Social Media, Cologne. (2016).
52. Das, S., and Kramer, A. Self-censorship on facebook. In: Proceedings of the International AAAI Conference on Web and Social Media, Boston, MA. (2013).
53. Wang, Y., Norcie, G., Komanduri, S., Acquisti, A., Leon, P. G., and Cranor, L. F. 'i regretted the minute i pressed share': A qualitative study of regrets on facebook. In: Proceedings of the Seventh Symposium on Usable Privacy and Security, SOUPS '11, New York, NY. 10:1–10:16 (2011).
54. Tasse, D., Liu, Z., Sciuto, A., and Hong, J. State of the geotags: Motivations and recent changes. In: Proceedings of the International AAAI Conference on Web and Social Media, Montreal, QC. (2017).

55. Hecht, B., and Stephens, M. A tale of cities: urban biases in volunteered geographic information. In: Proceedings of the International AAAI Conference on Web and Social Media, Ann Arbor, M. (2014).
56. Salganik, M. J. Bit by Bit: Social Research in the Digital Age. Princeton, NJ: Princeton University Press.(2017).
57. Lampe, C., Ellison, N. B., and Steinfield, C. Changes in use and perception of Facebook. In: Proceedings of the 2008 ACM Conference on Computer Supported CooperativeWork, CSCW'08. New York, NY. 721–730 (2008).
58. Liu, Y., Kliman-Silver, C., and Mislove, A. The tweets they are a-changin': Evolution of twitter users and behavior. In: Proceedings of the International AAAI Conference on Web and Social Media, Ann Arbor, MI. (2014).
59. Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., and Potts, C. No country for old members: user lifecycle and linguistic change in online communities. In: Proceedings of the 22nd International Conference on WorldWideWeb,WWW'13. New York, NY. 307–318 (2013).
60. Resnick, P., Garrett, R. K., Kriplean, T., Munson, S. A., and Stroud, N. J. Bursting your (filter) bubble: strategies for promoting diverse exposure. In: Proceedings of the 2013 Conference on Computer Supported Cooperative Work Companion, CSCW'13. New York, NY, 95–100 (2013).
61. Van Binh T., Minh D., Linh L., and Van Nhan T. Location-based service information disclosure on social networking sites: The effect of privacy calculus, subjective norms, trust, and cultural difference. Information Services & Use. 1-25. (2023).
62. Newell E., T., Dimitrov S., Piper A., and Van Ruths D. To Buy or to Read: How a Platform Shapes Reviewing Behavior. In: Proc. International Conference on Web and Social Media (ICWSM). (2021).
63. D'Alessio, D., and Allen, M. Media bias in presidential elections: a metaanalysis. J. Commun. **50** 133–156. (2000).
64. Blodgett, S. L., Green, L., and O'Connor, B. Demographic dialectal variation in social media: a case study of African-American English. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, 1119–1130 (2016).
65. Liang, H., and Fu, K.-w. Testing propositions derived from twitter studies: generalization and replication in computational social science. PLoS ONE 10:e0134270 (2015).
66. White, R. W. Interactions with Search Systems. Cambridge: Cambridge University Press. (2016).
67. Radford, J., & Joseph, K. Theory in, theory out: the uses of social theory in machine learning for social science. Frontiers in big Data, 3, 18. (2020).
68. Cerqueira, V., Torgo, L., Smailović, J., & Mozetič, I. A comparative study of performance estimation methods for time series forecasting. In 2017 IEEE international conference on data science and advanced analytics (DSAA) (pp. 529-538). IEEE, (2017, October).
69. Guest, E., Vidgen, B., Mittos, A., Sastry,N., Tyson, G., Margetts, H. An Expert Annotated Dataset for the Detection of Online Misogyny. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 1336–1350, Association for Computational Linguistics (2021).
70. Agarwal, P., Hawkins, O., Amaxopoulou, M., Dempsey, N., Sastry, N., Wood, E. Hate Speech in Political Discourse: A Case Study of UK MPs on Twitter. In: Proceedings of the 32nd ACM Conference on Hypertext and Social Media (HT '21). New York, NY, USA, 5–16 (2021).
71. Zia, H.B., Raman, A., Castro, I., Anaobi, I. H., Cristofaro, E.D., Sastry, N., Tyson, G. Toxicity in the Decentralized Web and the Potential for Model Sharing. In: Proc. ACM Meas. Anal. Comput. Syst. 6, 2, Article 35 (2022).
72. Vidgen, B., Thrush, T., Waseem, Z., Kiela, D. Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection. arXiv:2012.15761 (2021).
73. Yin, W., Agarwal, V., Jiang, A., Zubiaga, A., Sastry, N. AnnoBERT: Effectively Representing Multiple Annotators' Label Choices to Improve Hate Speech Detection. Accepted In: The 17th International AAAI Conference on Web and Social Media (ICWSM) (2023).