

# Fairness Auditing, Explanation and Debiasing in Linguistic Data and Language Models

Marta Marchiori Manerba<sup>1,2</sup>

<sup>1</sup>Computer Science Department, University of Pisa, Italy

<sup>2</sup>KDD Laboratory, ISTI, National Research Council, Pisa, Italy

## Abstract

This research proposal is framed in the interdisciplinary exploration of the socio-cultural implications that AI exerts on individual and groups. The focus concerns contexts where models can amplify discriminations through algorithmic biases, e.g., in recommendation and ranking systems or abusive language detection classifiers, and the debiasing of their automated decisions to become beneficial and just for everyone. To address these issues, the main objective of the proposed research project is to develop a framework to perform fairness auditing and debiasing of both classifiers and datasets, starting with, but not limited to, abusive language detection, thus broadening the approach toward other NLP tasks. Ultimately, by questioning the effectiveness of adjusting and debiasing existing resources, the project aims at developing truly inclusive, fair, and explainable models by design.

## Keywords

Responsible NLP, Explainability, Interpretability, Fairness

## 1. Introduction

At every stage of a supervised learning process, biases can arise and be introduced in the pipeline. Current models implemented with AI technologies have been shown to inherit and perpetuate bias against specific demographic groups and protected attributes such as sexual orientation or religion [1, 2]. These skews pose a severe risk and limitation to the well-being of underrepresented minorities, ultimately amplifying pre-existing social stereotypes, possible marginalization, and explicit harm [1, 3]. Given the sensitive contexts in which systems are deployed, a robust value-oriented evaluation of models' fairness is necessary to mitigate unfairness and avoid discrimination.

Besides fairness, another crucial aspect to consider lies in the opaqueness of models' internal behavior. If the dynamics leading a model to a particular automatic decision are not clear nor accountable, significant problems of trust for the reliability of outputs could emerge, especially in sensitive real-world contexts where high-stakes choices are made. Inspecting non-discrimination of decisions and assessing that the knowledge autonomously learned conforms to human values also constitutes a real challenge. Indeed, the objective of eXplainable Artificial Intelligence

---

*Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal*


✉ [marta.marchiori@phd.unipi.it](mailto:marta.marchiori@phd.unipi.it) (M. M. Manerba)

🌐 <https://martamarchiori.github.io/> (M. M. Manerba)

🆔 0000-0003-2251-1824 (M. M. Manerba)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

(XAI) is to propose strategies and methods to render AI systems and automatic decisions more intelligible to humans. In recent years, working towards transparency and interpretability of black box models has become a priority: multiple approaches and methods have been proposed [4, 5, 6].

This research topic can not be limited only to constructing mathematical explanations or those understandable only to data scientists, just as algorithmic fairness is not enough to effectively counteract certain types of harms [7]. Therefore, the phenomenon's complexity is not limited to algorithms but is deeply rooted and bound in historical, cultural, and social perceptions. It can not be solved by computational methods alone, nevertheless, what is guiding my commitment are the pressing need in demanding transparency and the intent of developing truly inclusive tools that can at least meet the needs of minorities' experiences in diverse social spaces.

To address these issues, the main objective of the proposed research project is to develop a framework to perform fairness auditing and debiasing of both classifiers and datasets, starting with, but not limited to, abusive language detection, thus broadening the approach toward other NLP tasks. The intuition relies on leveraging explainability techniques to discover biases and perform fairness auditing, e.g., generating counterfactuals and deploying interpretable proxies for the black box models. Auditing output can consist of an explanation of the (un)fairness of linguistic data and language models under analysis, i.e., specific reasons for which the resource is considered unfair, ultimately exposing unjust behaviors more visibly and transparently. The framework will propose several metrics and strategies to quantify and approach debiasing from the identified discriminatory treatments, beginning with those attested within ML and NLP communities. Ultimately, by contesting the effectiveness of adjusting and debiasing existing resources, the project aims at developing truly inclusive, fair, and explainable models by design. In Sections 3 and 4, we will describe in detail the proposed strategies.

## 2. Related Work

**XAI and Fairness Approaches for ML and NLP.** Overall, few approaches in the literature are at the intersection of explainability and fairness. The use of XAI techniques to identify and explain fairness issues is presented in [8]. The authors, highlighting the gap in this direction of research, outline generic recommendations for devising XAI tools, specifically proposing guidelines for the development of a *Fair Explainability toolkit*, after outlining the steps in the design, planning, deployment, and use of AI systems in which bias can potentially be introduced. This toolkit should be able to: (1) investigate the source data, (2) highlight the impacts of the choice and development of ML models, and (3) design explanations according to the identified target audience. One branch emerging from this intersection is the assessment of the fairness of the explanations by checking the fidelity score of the explanations calculated for each sensitive group [9]. The investigation is carried out to evaluate if the explanations are more suitable for a specific group w.r.t. others. However, this type of assessment is highly dependent on the choice of explainer and how the hyperparameters are set: the unreliable nature, i.e., as other explainers produce different explanations for the same instances, constitutes a significant challenge. Another dimension concerns the study of how explanations impact users' perceptions of fairness if, indeed, they can increase human trust in the fairness and correctness of automatic

decision-making [10]. We refer to the review conducted in [11], where authors collect works that propose strategies to tackle the fairness of NLP models through explainability techniques. Generally, authors found that, although one of the main reasons for applying explainability to NLP resides in bias detection, contributions at the intersection of these ethical AI principles are very few and often limited in the scope, e.g., w.r.t. biases and tasks addressed. Additionally, when considering the integration of fairness and XAI, it is essential to recognize the distinct objectives of each. Fairness primarily emphasizes equitable outcomes, whereas XAI concentrates on enhancing transparency and understanding of the underlying processes. As the authors point out [11], there is a lack of metrics to address procedural fairness. Therefore, such approaches, which apply XAI to fairness questions, are often reduced to checking that sensitive attributes are not used in decisions, ultimately implementing the much more problematic approach of fairness through unawareness, which attempts to achieve fairness by intentionally ignoring or not considering sensitive attributes during decision-making processes. Indeed, fairness through unawareness has been criticized for several reasons. First, it assumes that excluding sensitive attributes automatically eliminates bias, disregarding the potential influence of other correlated attributes that can still perpetuate discriminatory outcomes. Second, it overlooks that ignoring sensitive attributes can hinder the identification and understanding of discriminatory patterns and potential biases in the system. Consequently, fairness through unawareness can mask underlying biases and hinder the ability to address and rectify unfairness effectively. Conversely, fairness through awareness [12] acknowledges the existence of such attributes and takes into explicit account the potential impacts on different groups. It aims to address biases and ensure equitable outcomes by actively recognizing and mitigating disparities associated with these attributes since they can be relevant and important factors in certain contexts.

**Challenges.** We follow the insights from the review conducted in [13], where authors provide an overview of the current state of XAI and its relationship to NLP. Currently, explainability approaches generally work on low-dimensional tabular data and take a long time to run, so they do not scale to other types or large-scale datasets. Most explanation methods for NLP applications are local, remain at the analysis of the linguistic surface, and therefore expose mostly non-causal relations. Regarding the limitations of leveraging XAI to improve the fairness of NLP models[11], if the explanation methods are not robust, consistent, and thus reliable, using explanations to delegate or certify shallow fairness is a risk. Moreover, it is crucial to introduce the concept of the “uncertainty level” of the explanation to help the user understand how much it is possible to rely on the explanation. Both explainability and fairness face the challenge of lacking shared terminology and recognized standards, as there is still no full agreement within the field. This lack of consensus arises from the diverse range of datasets and models used, making it difficult to establish consistent frameworks for systematic comparisons and benchmarking. Although it is a priority to raise new and complex questions within human-centered ML, assessing the impacts on individuals and understanding what users count as fair, human-in-the-loop has its costs. Nevertheless, the opportunity to conduct robust user testing would be essential, collecting human evaluation and fairness judgments to improve the suitability and quality of explanations w.r.t. specific contexts. Although these challenges are extremely limiting to the pursuit of developing fair and explainable NLP techniques that are also robust and reliable, I believe these limitations can be a starting point for my project, through which I intend to overcome some of them by addressing them together in a systemic,

participatory, co-design and continuous correction perspective, to include missing, unheard voices and sensitivities, “*interrogating and reimagining the power relations between technologists and such communities*” [14].

### 3. Research Questions and Approach

Leveraging XAI and interpretability strategies to uncover fairness issues [15], we intend to address this problem by designing a framework that deals with detection and mitigation aspects. Defining solutions to address the unfairness requires considering various dimensions, such as what constitutes a sensitive attribute to be protected or according to which criteria it is possible to assess whether a decision is fair.

1. **Can explainability techniques contribute to discovering the source and the reason for unfair, biased behaviors in NLP pipelines?**
  - a) *Which explainability techniques help most to uncover biases in NLP applications?*
  - b) *What consists of a meaningful explanation for NLP applications w.r.t. the developers in order to expose potential harm?*
  - c) *What about the essential features of an explanation addressed to the final users, enabling them to both understand the reasons behind automatic decisions and to appeal for recourse?*
2. **How to implement explainable and fair by design approaches?**

Contributions at the intersection of these fields are still at the start, as reported by [11], where trends in XAI and fairness in NLP research are reviewed. Current solutions are restricted to a few tasks, address narrow biases, and leverage mainly local explanation methods. Since fairness and explainability are young disciplines and lack solid theoretical foundations, collaboratively building at the intersection of these two AI ethics principles might be a promising strategy for exposing the bias. I want this work to position itself differently from the existing literature, starting with the clear articulation -still under development- of the concepts we want to deal with, i.e., bias and unfairness in NLP, to identify what to measure and mitigate effectively. Using explainability to uncover fairness issues is instead motivated by the lack of transparency. The inability to provide explanations for AI systems is also often blamed as a source of bias [11]. Explainability, in this sense, becomes an analytical tool to shed light on both the outputs and internal dynamics of systems to identify and motivate unjust automatic behaviors. This contamination within NLP is so far potential and underinvestigated, as very few (and insufficient) approaches have explored and devised solutions at the intersection of fairness and explainability, as reported in [11]. As for using post-hoc explainability to generate explanations of fairness for model behavior, the intuition might be to build methods that produce explanations with linguistic structure in consideration, exploiting it to account for implicit and less superficial language dynamics, i.e., going beyond the counterfactual token fairness metric. Regarding datasets, the goal could be to design effective guarantees that manage to train a model and issue fair decisions even in the presence of biased data.

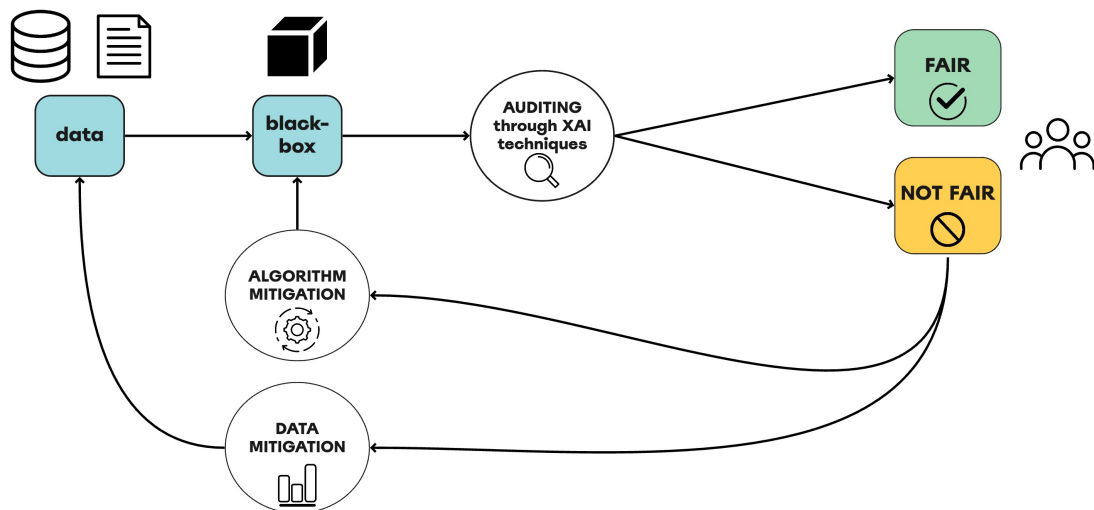


Figure 1: Fairness Evaluation Loop workflow.

## 4. Research Directions and Next Steps

The following section presents potential research lines and solutions to the questions under investigation. The concrete aim of this research is intended to address, on the one hand, the assessment and “adjustment” of existing tools through a strong value-oriented evaluation [16]. On the other hand, the urgent need to develop truly inclusive tools, fair and explainable by design.

**Resources to counteract stereotypes and infer fairness explanations.** We report a workflow hypothesis of a *Fairness Evaluation Loop* (in Fig. 1 a visual representation). Studying the interplay between XAI, fairness, and ethics, it aims at unmasking, detecting, and counteracting bias within NLP applications, ultimately fostering responsible ML. The framework, pursuing fairness as a multi-objective task, will combine:

- evaluation/detection: risk assessment approaches exploiting, among others, (1) in-depth analysis of the performances/errors obtained over demographic groups; (2) XAI techniques, e.g., the generation of counterfactuals and the deployment of interpretable proxies; (3) other ML techniques, such as the detection of outliers predictions;
- debiasing/mitigation of data, keeping the algorithm fixed but retraining it;
- debiasing/mitigation of the algorithm or fair algorithm development, with data for which we can not guarantee;
- continuous cycle of monitoring and correction.

The output will consist of different explanations according to different user types.

**Resources that are responsible by design, meaning fair and explainable.** Despite the importance of mitigating unfairness and explaining opaque systems, the challenges and limitations of current approaches demonstrate how complex and multifaceted the task is and how occasionally, instead of solving the problem, others are introduced. A promising research direction, beyond debiasing and explainability, could concern the development of truly inclusive models, fair and explainable by design regardless of the potential bias in the data [17, 18]. One contribution could be the collection and publication of representative datasets containing instances of the misrepresented [15] phenomena. It is crucial to assess and address especially the under-recognised ones since biases are manifestations of distinct stereotypes and not all have received the same attention from the scientific community so far. Another promising line, justified by the need for users' acceptance, could regard the design of participatory approaches at different involvement levels and stages of the ML pipeline for detecting risks and harms. Certain biases require the engagement and the feedbacks of the affected groups to be effectively exposed and addressed [15].

To evaluate both lines of research, we will conduct several experiments in order to demonstrate the novelty of our approach compared to other SOTA explainers and bias-assessment benchmark procedures, proving the effectiveness of our framework in unmasking biases in both research and commercial systems as well as the main benchmarks and gold standard datasets for several NLP tasks of interest, starting with, but not limited to, abusive language detection. Since state of the art techniques for dealing with fairness operate mainly on tabular data, we will build on existing techniques by expanding the approaches toward NLP applications and models that operate on textual data.

As suggested in [16], proposing a contribution within the NLP domain responsibly and consciously means foremost acknowledging our own biases. This might mean starting by recognising how most contributions currently reflect a dominant perspective and culture, thus unconsciously incorporating stereotypes and marginalisation. Furthermore, it is crucial to overcome the techno-solutionism, being aware that any solely technological solution will be partial, as not considering the broader socio-political issue that is the source of these biases means simplifying and “fixing” only on the surface [7]. We must remember that “resolving the bias” does not guarantee the ethical use of technology. A systemic approach is necessary, combined with creating a narrative that avoids misrepresenting and mystifying these complex socio-technical tools. Regardless, we firmly believe that NLP pipelines need a robust value-sensitive evaluation in order to assess unintended biases and avoid, as far as possible, explicit harm or the amplification of pre-existing social prejudices, trying to ultimately build systems that contribute in a beneficial way to the society and all its citizens.

## Acknowledgments

This work has been partially supported by the European Community Horizon 2020 programme under the funding scheme ERC-2018-ADG G.A. 834756 *XAI: Science and technology for the eXplanation of AI decision making*.



## References

- [1] H. Suresh, J. V. Guttag, A framework for understanding unintended consequences of machine learning, *CoRR abs/1901.10002* (2019).
- [2] N. Mehrabi, et al., A survey on bias and fairness in machine learning, *ACM Comput. Surv.* 54 (2021) 115:1–115:35.
- [3] L. Dixon, J. Li, J. Sorensen, N. Thain, L. Vasserman, Measuring and mitigating unintended bias in text classification, in: *AIES, ACM*, 2018, pp. 67–73.
- [4] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, *ACM Comput. Surv.* 51 (2019) 93:1–93:42.
- [5] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), *IEEE Access* 6 (2018) 52138–52160.
- [6] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, *Artif. Intell.* 267 (2019) 1–38.
- [7] E. Ntoutsi, et al., Bias in data-driven artificial intelligence systems - an introductory survey, *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 10 (2020).
- [8] K. Alikhademi, B. Richardson, E. Drobina, J. E. Gilbert, Can explainable AI explain unfairness? A framework for evaluating explainable AI, *CoRR abs/2106.07483* (2021). URL: <https://arxiv.org/abs/2106.07483>. arXiv: 2106.07483.
- [9] A. Balagopalan, H. Zhang, K. Hamidieh, T. Hartvigsen, F. Rudzicz, M. Ghassemi, The road to explainability is paved with bias: Measuring the fairness of explanations, *arXiv preprint arXiv:2205.03295* (2022).
- [10] K. Orphanou, J. Otterbacher, S. Kleanthous, K. Batsuren, F. Giunchiglia, V. Bogina, A. S. Tal, A. Hartman, T. Kuflik, Mitigating bias in algorithmic systems-a fish-eye view, *ACM Computing Surveys (CSUR)* (2021).
- [11] E. Balkir, S. Kiritchenko, I. Nejadgholi, K. C. Fraser, Challenges in applying explainability methods to improve the fairness of nlp models, *arXiv preprint arXiv:2206.03945* (2022).
- [12] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. S. Zemel, Fairness through awareness, in: S. Goldwasser (Ed.), *Innovations in Theoretical Computer Science 2012*, Cambridge, MA, USA, January 8-10, 2012, ACM, 2012, pp. 214–226. URL: <https://doi.org/10.1145/2090236.2090255>. doi:10.1145/2090236.2090255.
- [13] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, P. Sen, A survey of the state of explainable AI for natural language processing, in: K. Wong, K. Knight, H. Wu (Eds.), *AAACL/IJCNLP 2020*, Suzhou, China, December 4-7, 2020, Association for Computational Linguistics, 2020, pp. 447–459. URL: <https://aclanthology.org/2020.aacl-main.46/>.
- [14] S. L. Blodgett, S. Barocas, H. D. III, H. M. Wallach, Language (technology) is power: A critical survey of "bias" in NLP, in: D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.), *ACL 2020*, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 5454–5476. URL: <https://doi.org/10.18653/v1/2020.acl-main.485>. doi:10.18653/v1/2020.acl-main.485.
- [15] L. Weidinger, et al., Taxonomy of risks posed by language models, in: *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 214–229.
- [16] R. Dobbe, S. Dean, T. K. Gilbert, N. Kohli, A broader view on bias in automated decision-making: Reflecting on epistemology and dynamics, *CoRR abs/1807.00553* (2018).

- [17] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nature Machine Intelligence* 1 (2019) 206–215.
- [18] C. Wang, B. Han, B. Patel, F. Mohideen, C. Rudin, In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction, *CoRR* abs/2005.04176 (2020).