

Eliciting Discrimination Risks in Algorithmic Systems: Taxonomies and Recommendations

Marta Marchiori Manerba¹

¹Università di Pisa, ISTI-CNR (Pisa, Italy)

Abstract

The main objective of this preliminary study is to account for how discrimination is interpreted, encoded, and addressed in contributions proposing categorizations of risks and harms emerging from AI systems and Language Models in particular. Therefore, we delve into surveys and reviews that frame algorithmic harms, providing a preliminary overview of strategies and alternative perspectives that are not purely technical. We conclude by promoting the positive role of these contributions while simultaneously highlighting their inherent limitations and the caution needed to implement such broad and far-reaching recommendations in practice.

Keywords

Sociotechnical Risks, Algorithmic Harms, Discrimination, Fairness, Bias, Stereotypes

Introduction

Current models have been shown to inherit and perpetuate bias towards 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]. To explore these issues, the main objective of this preliminary study is to account for how discrimination is interpreted, encoded, and addressed in contributions proposing categorizations of risks emerging from AI systems. Progressively narrowing the scope of the investigation, we specifically focus on how unfairness can be recognized and mitigated in Language Models (LMs). We, therefore, gather recommendations and alternative practices from various actors and disciplines to broaden the discussion beyond purely technical solutions.

Taxonomies of Risks: Identifying Bias

In the following, we dig into surveys and literature reviews that frame algorithmic harms. We select contributions from the most recent and adopted works, transitioning from a broad framing of AI to Foundation Models and LMs. Several risks are indeed common and cross-cutting across many AI systems, while new dangers are more specific to the peculiarities of LMs, especially

EWAF'24: European Workshop on Algorithmic Fairness, July 01–03, 2024, Mainz, Germany


✉ marta.marchiori@phd.unipi.it (M. M. Manerba)

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

🆔 0000-0003-2251-1824 (M. M. Manerba)



© 2022 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)

given the latest technological advancements that have brought them to the forefront and made them urgent to take into account and address.

Shelby et al. [4] propose a taxonomy encompassing sociotechnical harms arising from algorithmic systems and related prevention and mitigation strategies. This categorization is general and looks at the impacts that manifest as a result of the use of technology in, on, and by society, focusing on consequences that propagate in real-world contexts. Indeed, *sociotechnical* acknowledges that harm recognition must be conducted in real-world application contexts, where a nuanced interplay occurs between the technology and the socio-cultural tissue featuring complex power dynamics. Unsurprisingly, marginalized communities suffer this kind of damage the most, reflecting being historically targeted w.r.t. discrimination and social exclusion: in fact, the technology often reproduces and reinforces default social power patterns, embedding unjust worldviews. In the following list, we report the categorization proposed by the authors. As they highlight, buckets are not mutually exclusive since harms often arise concurrently and multiply: (1) Representation Harms; (2) Allocative Harms; (3) Quality of Service Harms; (4) Interpersonal Harms; (5) Social System Harms.

In Bommasani et al. [5], an in-depth review of so-called *foundation models* is outlined, spanning from technical functioning, abilities, applications and social impact, which authors state is complex to grasp since these systems have not yet specific purposes, but rather serve as a building block, and in so, reasoning on risks become even more challenging, although necessary especially due to the fact that the abilities of LMs are mainly based on the foundations, that heavily affects the properties and subsequent models behaviours. The broad risks examined include *fairness, unintended uses, e.g., misinformation, homogenization, impacts on environment, global economy and politics, and existing legislative regulations*. Here, we focus on harms pertaining to fairness since the report is extensive, as briefly outlined, covering various aspects and delving into numerous topics, not all relevant to the scope of our framing. First, authors introduce the concepts of intrinsic, latent biases, i.e., “*properties of the foundation model can lead to harm in downstream systems*”, and extrinsic harms, i.e., “*specific harms from the downstream applications that are created by adapting a foundation model*” of which users experience. Representational bias pertains to the intrinsic harms and manifests itself as misrepresentation, underrepresentation, and overrepresentation. At the extrinsic level, representational bias manifests in the generation of abusive content, and marked performance disparities among different demographic groups. A crucial prerequisite for an ethical assessment of these systems lies in building a shared awareness and understanding of groups and prejudices. Training and adaptation data (especially when not fully known/scrutinizable), modelling, modeler diversity, and community values are identified as potential bias sources.

We now shift the lens towards the risks of harms arising from Generative LMs framed by Weidinger et al. [6, 7]. Authors identify six risk areas that overlap with and reference the previously discussed taxonomy [4], with the difference that this categorization is more fine-grained for the NLP scenario, providing a more detailed description of the dangerous manifestations originating from the identified harms. The six areas are: (1) Discrimination, Hate speech and Exclusion; (2) Information Hazards; (3) Misinformation Harms; (4) Malicious Uses; (5) Human-Computer Interaction Harms; (6) Environmental and Socioeconomic Harms. We focus on the discrimination arguments at the core of this review, i.e., where “*the LM accurately reflects unjust, toxic, and oppressive speech present in the training data*”. First, it is crucial to

acknowledge that the emergence of these risks “originates” from the LMs’ emulation of natural language, which embeds unfair, abusive, and unbalanced power dynamics ingrained within its training data.

Cui et al. [8] formalize a risk taxonomy for LLMs broken down with respect to the various model components. Compared to previous taxonomies, it situates and places the risks based on where, in which part, and at what point of the entire pipeline they manifest. We only highlight the risks related to discrimination: (1) Toxicity and Bias Tendencies (LM module): “*extensive data collection in LLMs brings toxic content and stereotypical bias into the training data*”; (2) Harmful Content (Output module): “*the LLM-generated content sometimes contains biased, toxic, and private information*”.

From the overview outlined, an evident overlap emerges between the risks and harms encoded in various taxonomies. Authors often refer to similar concepts using different expressions. This overlapping is certainly positive as it demonstrates a sharing of perspectives and priorities, indicating agreement and a collective effort that the community is undertaking to unify this research space. Establishing a priori, in abstract, which taxonomy and formulation is most suitable is impossible: each application context will correspond to a subset of risks that are more relevant and impactful than others. Certain categories of harms will therefore be more prominent for specific applications, while they may be more negligible and less central for others, where they might be considered as additional, potential impacts. On the other hand, however, having a multitude of contributions addressing and formalizing the same categories could potentially lead to increased confusion. A consequence, certainly unintended but severe, consists of the so-called “ethical shopping”, where developers and providers (both public and private) may choose the most convenient lens that best suits their technology without having to overhaul it to mitigate the identified risks, promoting a sterile and self-serving practice [9]. To wrap up this overlook, we emphasize how defining and operationalize fairness in the NLP context is challenging. A severe lack of consensus persists in framing undesirable outcomes—such as bias, fairness, justice, and harms—raising the question of what might change when we alter our framings. Existing works are often inaccurate, inconsistent, and contradictory in formalizing bias, as demonstrated by Blodgett et al. [10]. Clarifying the concepts related to harms arising from model behaviors, often underspecified and overlapping, is a prerequisite for proposing a contribution that explores and embraces the link between language, technology, and social structures. Behind every implementation choice, the implicit set of social values and structures that justifies and grounds the technical solution proposed must not be left implied.

Beyond Technical Lens: Development Guidelines and Governance Approaches

Feminist and critical race theories from the human-computer interaction field offer diverse perspectives and standpoints to embrace and experiment within AI design, offering both critique-based and generative, original solutions. Specifically, adopting the feminist lens, values, and practices to be embedded in AI processes are: *pluralism, participation, advocacy, ecology, embodiment, and self-disclosure* [11]. In [12], research directions are outlined from the integration of feminist stances into explainability, giving importance to disregarded minorities’ voices as a different option to the universalist “one explanation fits all” default stance and truly engaging in diversity and structural power dynamics. The feminist intersectional principles outlined for

XAI include (i) *normative orientation towards social justice and equity*, (ii) *attention to power and structural inequalities*, (iii) *challenging traditional rationalist modalities of explanation*, and (iv) *centring marginalised perspectives*. From adopting these values in the design, implementation, and deployment of (X)AI systems, practical implications are drawn, such as context-dependence, the concept of proactivity, and aiming at participation and interaction as fundamental properties of the explanations. State and Fahimi [13] propose to *carefully* examine explanations assessing existing techniques from a feminist viewpoint. Authors also introduce the notion of *caring XAI*, i.e., rethinking about explanations as a *caring practice*.

From critical race theory, a mature, serious acknowledgment of the ubiquitousness of racism in pipelines is referred to as the first crucial step [14]. This awareness must guide the community in reflexive attitudes, starting from critically assessing the representativeness of data and teams, re-framing narratives so that they can be accessible and meaningful not only to the racial majority, broadening works' impacts assessment towards the racial axis, including race-sensitive contributions, organizing open panels, workshops, public discussion and producing collaborative statements and initiatives to raise and promote shared concerns and care towards these societal phenomena. Marda and Narayan [15] state the crucial role of qualitative ethnographic methodologies in the AI field as a means to thoroughly comprehend the social impact of these technologies, delving into the power relationships between actors, subjects, and institutions. Specifically, highlighting the inherent limitations of quantitative methods, ethnographic stances aid researchers and the community in answering critical questions related to the *how* and *why* of algorithmic results, investigating and questioning also the actors that hold unbalanced power dynamics and constructively exposing assumptions taken for granted [10]. Likewise, to reconceptualize AI ethics beyond Western lens, which often emphasise the priority of preserving the rights of individuals instead of promoting the welfare of the community as a whole, Amugongo et al. [16] propose to leverage the African philosophy of Ubuntu. The principles emerging from Ubuntu, focusing on interconnectedness and interdependence of the collectivity, are *fairness*, *community good*, *safeguarding humanity*, *respect for others*, and *trust*.

Outlook

From these contributions, it is evident that socio-technical lenses are required to deal holistically with the societal problems that AI fosters and amplifies and that techno-solutionist positivist uncritical approaches on their own are not enough to tackle the roots of these pervasive risks. Indeed, it is crucial to understand that any solely technological solution will be partial. Not considering the broader socio-political issue that is the source of these biases means simplifying and “fixing” only on the surface [17]. Finally, we must remember that “resolving the bias” does not guarantee the ethical use of technology, a direction beyond the scope of this work. A systemic approach is necessary, combined with creating a narrative that avoids misrepresenting and mystifying these complex, albeit controllable, socio-technical tools. It is essential to remember that technology itself is (almost always) not inherently bad, on its own: it is the specific human adoption that results in beneficial or harmful consequences. An open and urgent area is developing these technologies so that the design is explicitly oriented towards and promotes ethics. To this end, interdisciplinary research and development teams are necessary to address the issue holistically.

Acknowledgments

This work has been 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, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, 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. Shelby, S. Rismani, K. Henne, A. Moon, N. Rostamzadeh, P. Nicholas, N. Yilla-Akbari, J. Gallegos, A. Smart, E. García, G. Virk, Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction, in: F. Rossi, S. Das, J. Davis, K. Firth-Butterfield, A. John (Eds.), *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2023, Montréal, QC, Canada, August 8-10, 2023*, ACM, 2023, pp. 723–741. URL: <https://doi.org/10.1145/3600211.3604673>. doi:10.1145/3600211.3604673.
- [5] R. Bommasani, D. A. Hudson, E. Adeli, R. B. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. S. Chatterji, A. S. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. D. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. S. Krass, R. Krishna, R. Kuditipudi, et al., On the opportunities and risks of foundation models, CoRR abs/2108.07258 (2021). URL: <https://arxiv.org/abs/2108.07258>. arXiv:2108.07258.
- [6] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh, et al., Taxonomy of risks posed by language models, in: *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 214–229.
- [7] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, Z. Kenton, S. Brown, W. Hawkins, T. Stepleton, C. Biles, A. Birhane, J. Haas, L. Rimell, L. A. Hendricks, W. Isaac, S. Legassick, G. Irving, I. Gabriel, Ethical and social risks of harm from language models, CoRR abs/2112.04359 (2021). URL: <https://arxiv.org/abs/2112.04359>. arXiv:2112.04359.
- [8] T. Cui, Y. Wang, C. Fu, Y. Xiao, S. Li, X. Deng, Y. Liu, Q. Zhang, Z. Qiu, P. Li, Z. Tan, J. Xiong, X. Kong, Z. Wen, K. Xu, Q. Li, Risk taxonomy, mitigation, and assessment benchmarks of large language model systems, CoRR abs/2401.05778 (2024). URL: <https://doi.org/10.48550/arXiv.2401.05778>. doi:10.48550/ARXIV.2401.05778. arXiv:2401.05778.
- [9] L. Floridi, *The ethics of artificial intelligence: principles, challenges, and opportunities* (2023).

- [10] 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. Schlueter, J. R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 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.
- [11] S. Bardzell, Feminist HCI: taking stock and outlining an agenda for design, in: E. D. Mynatt, D. Schoner, G. Fitzpatrick, S. E. Hudson, W. K. Edwards, T. Rodden (Eds.), Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, Atlanta, Georgia, USA, April 10-15, 2010, ACM, 2010, pp. 1301–1310. URL: <https://doi.org/10.1145/1753326.1753521>. doi:10.1145/1753326.1753521.
- [12] G. Klumbyte, H. Piehl, C. Draude, Towards feminist intersectional XAI: from explainability to response-ability, CoRR abs/2305.03375 (2023). URL: <https://doi.org/10.48550/arXiv.2305.03375>. doi:10.48550/ARXIV.2305.03375. arXiv:2305.03375.
- [13] L. State, M. Fahimi, Careful explanations: A feminist perspective on XAI, in: J. M. Alvarez, A. Fabris, C. Heitz, C. Hertweck, M. Loi, M. Zehlike (Eds.), Proceedings of the 2nd European Workshop on Algorithmic Fairness, Winterthur, Switzerland, June 7th to 9th, 2023, volume 3442 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: <https://ceur-ws.org/Vol-3442/paper-39.pdf>.
- [14] I. F. Ogbonnaya-Ogburu, A. D. R. Smith, A. To, K. Toyama, Critical race theory for HCI, in: R. Bernhaupt, F. F. Mueller, D. Verweij, J. Andres, J. McGrenere, A. Cockburn, I. Avellino, A. Goguey, P. Bjøn, S. Zhao, B. P. Samson, R. Kocielnik (Eds.), CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020, ACM, 2020, pp. 1–16. URL: <https://doi.org/10.1145/3313831.3376392>. doi:10.1145/3313831.3376392.
- [15] V. Marda, S. Narayan, On the importance of ethnographic methods in AI research, *Nat. Mach. Intell.* 3 (2021) 187–189. URL: <https://doi.org/10.1038/s42256-021-00323-0>. doi:10.1038/s42256-021-00323-0.
- [16] L. M. Amugongo, N. J. Bidwell, C. C. Corrigan, Invigorating ubuntu ethics in AI for healthcare: Enabling equitable care, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023, ACM, 2023, pp. 583–592. URL: <https://doi.org/10.1145/3593013.3594024>. doi:10.1145/3593013.3594024.
- [17] E. Ntoutsis, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, I. Kompatsiaris, K. Kinder-Kurlanda, C. Wagner, F. Karimi, M. Fernández, H. Alani, B. Berendt, T. Kruegel, C. Heinze, K. Broelemann, G. Kasneci, T. Tiropanis, S. Staab, Bias in data-driven artificial intelligence systems - an introductory survey, *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 10 (2020).