Formal ontologies for fair language technologies: A conceptual exploration

Annapaola Ginammi[1]

[1]META, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy

Abstract

Natural Language Processing (NLP) systems such as chatbots, resume screening systems, and machine translation tools tend to reflect societal biases such as gender bias, and mitigating these biases is widely regarded as crucial for ethical use of these language technologies. Mitigating bias is not merely a technical challenge, but requires distinguishing legitimate semantic content from bias, a task that involves both conceptual and normative commitments. Focusing on gender bias mitigation, this paper explores, from a conceptual point of view, the potential of formal ontologies (FOs) as a framework for making these commitments explicit. It presents FOs as an interdisciplinary methodology that can explicitly represent normative conceptions of what is taken to be legitimate semantic content underlying NLP bias mitigation approaches, thereby allowing assumptions that usually remain implicit to be systematically analyzed and evaluated.

Keywords

Natural Language Processing, Formal Ontology, social bias, normativity

1. Introduction

If our goal is to develop language technologies that benefit all equally, that is, that do not systematically lead to more positive outcomes for some individuals or groups of people than for others, then bias is a major concern. Bias, in NLP systems, may be understood broadly as the set of factors that contribute to unfair outcomes, such as allocative and representational harms [1]. NLP systems have been shown to induce such harms, for example by disadvantaging women in hiring decisions when a resume screening tool associates being an engineer with being a man, or by reinforcing gender stereotypes when an LLM systematically labels men as ambitious and confident, and women as collaborative and helpful in a performance evaluation [2]. Fairness, in this context, may then be understood as the avoidance of these harms. Progress towards fair NLP systems has been made but problems persist [2, 3, 4, 5], calling for a deeper understanding of how current interventions relate to the aims and underlying assumptions of NLP bias mitigation.

Formulating the aims of NLP bias mitigation is complex from both a conceptual and ethical perspective. This complexity arises because multiple conceptions of fairness coexist, these conceptions must be operationalized within probabilistic models, and natural language is itself inherently shaped by social dynamics [1, 3, 6, 7, 8, 9]. As a result, it is difficult to articulate a unified and well-justified objective for NLP bias mitigation, that is, a coherent and sufficiently detailed conception of what constitutes a "fair" NLP system. In fact, existing interventions pursue different, and sometimes conflicting, notions of fairness or neutrality, some of which rest on questionable methodological assumptions, and claims that an intervention "significantly reduces" or even "eliminates" bias are necessarily relative to the particular measure of bias being evaluated [9].

At the same time, notions of fairness or neutrality are often implicitly assumed by bias mitigation approaches. Bias is typically treated as a distortion that can be removed while preserving an "unbiased" or "neutral" semantic core of the model, where this semantic core is rarely made explicit. For example, removing the gender associations of "computer programmer" and "home maker" [10], implicitly

BEWARE-26 (Bias, Risk, Explainability and the role of Logic and Logic Programming), 8 October 2026, Perugia, Italy, co-located with AIxIA 2026.

annapaolaginammi@gmail.com (A. Ginammi)
https://annapagina.wordpress.com (A. Ginammi)
0000-0001-7153-3419 (A. Ginammi)

© 2026 Copyright for this paper by its author. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-17

CEUR-WS.org/Vol-4332/paper7.pdf

presupposes a semantic baseline according to which these occupational terms are gender neutral. Yet this purportedly neutral semantic baseline is itself normative: rather than a value-free description of linguistic meaning, it reflects a particular conception of how language ought to represent social categories. Weakening gender associations of occupational terms does not amount to removing ("bad", or "outdated") norms, but rather to encoding other ("good" or "contemporary") ones. Leaving these normative commitments implicit obscures the value judgments that guide bias mitigation, and it will be the norms of the NLP community that are effectively encoded in NLP systems [1].

This paper explores FOs as a framework for addressing both issues. FOs are logic-based specifications of the concepts and relations structuring a domain [11], and as such offer a principled way of specifying the semantic commitments implicit in existing bias mitigation approaches. These semantic commitments may be specified according to a chosen conception of fairness, and as such FOs may make fairness ideals concrete and open to critical examination. FOs in which fairness ideals are operationalized may foster interdisciplinary and intersectional analysis and debate, and function as a target against which bias mitigation techniques can be evaluated. The present paper offers conceptual reflections that are to be further investigated and tested in future research. In line with the bulk of NLP bias research, its main focus will be gender bias and associations between gender and occupational terms.

The paper is set up as follows. Section 2 argues that NLP bias mitigation requires explicit normative commitments. Section 3 proposes formal ontologies as a framework for representing these commitments. Section 4 argues that this framework also facilitates interdisciplinary collaboration. The paper concludes by outlining directions for future research.

2. The normative component of NLP bias mitigation

Any NLP bias mitigation procedure needs to distinguish the learned regularities that should remain from those that should be altered. It needs to distinguish, in other words, which of these regularities constitute legitimate semantic content[1] and which, instead, are regarded as distortions of that content. For example, when Bolukbasi et al. remove the gender direction from "computer programmer" [10], they are committed to something along these lines: there is a legitimate representation of the meaning of "computer programmer" of which gender is not a constitutive semantic feature.

Now, the statistical structure of a language model by itself does not provide a criterion that can ground a distinction between legitimate semantic representations and distortions or "bias". From the perspective of distributional statistics alone, a strong association between "computer programmer" and "man" is no different from that between "Rome" and "Italian". Thus, NLP bias mitigation must appeal to criteria external to the statistical structure of the model to distinguish which learned patterns are legitimate targets of intervention. Importantly, neither is linguistics sufficient to ground such a distinction. For example, overtly gendered lexical forms are sometimes treated as encoding legitimate semantic information and are therefore exempted from debiasing (e.g., in [10, 14]), but in other cases they are criticized and replaced by gender-neutral alternatives (consider "mankind", "chairman"; [15, 16, 3]). Thus, linguistic form does not by itself determine whether a gender distinction ought to be preserved or eliminated. The distinction between legitimate semantic content and bias is therefore neither statistical nor purely linguistic, and rather contains a substantial normative component.

The normative component of NLP bias research generally remains implicit and obscured by technical methodology. For example, one of the most commonly used frameworks for gender bias detection in NLP applications is the Word Embedding Association Test (WEAT), which was initially derived from psychology's Implicit Associations Test (IAT) and adapted to word embeddings [17], and later further adapted in various ways to LLMs and computer vision [5]. WEAT quantifies how strongly two sets of target words (e.g. male/female terms) are associated with two sets of attribute words (e.g.

[1] Bias mitigation in language models seems to fit particularly well with negotiation-based accounts of semantic content, which emphasize that meaning is shaped, at least in part, through (more or less tacit) negotiation over how best to use words [12, 13]. From this perspective, NLP bias mitigation can be understood as participating in an ongoing process of negotiating what words should mean and how they should be used.

career/family) based on the cosine similarity of their vector representations. However, qualifying a stronger association between, e.g., "woman" and "home" than "man" and "home" as biased presupposes normative assumptions about social fairness. WEAT thus uses technical methods but relies on normative assumptions in interpreting results.

Making the normative assumptions of bias mitigation explicit is important also for understanding whether or not bias mitigation serves the aim of developing language technologies that benefit all equally. This is not always as straightforward as it might seem. For example, in task-specific bias evaluation in LLMs, Gallegos et al. substituted gender in a question-answering task to understand if and how the response from the model changes [3]. To the question "If I am a male data engineer, what would you recommend I focus on?" the system responded: "We recommend focusing on identifying the applicable data sources", whereas to "If I am a female data engineer, what would you recommend I focus on?" the system responded instead: "As a female data engineer, it's important to focus on developing strong technical skills". Bias mitigation might be justified in this case on the ground of the system's reliance on the stereotype of females being less competent than males. However, in fact, female data engineers might encounter a more hostile environment than male data engineers, and therefore a focus on developing strong technical skills might be beneficial to them. It is an open question whether or not language technologies should keep track of such facts. Patterns like these are clearly gendered and socially meaningful, but whether or not they should be removed if our goal is to make language technologies fairer is contestable. The answer to such questions is far from obvious, and for sure they are neither technical nor purely linguistic.

Neither are such questions straightforwardly answered by the adoption of a certain conception of fairness. For example, adopting the fairness principle of treating like cases alike [6], leaves undecided whether an NLP system in a question-answering task should provide the same recommendations to male as to female data engineers. Deciding this requires reflection on to which extent, and under which conditions, male data engineers are "like" female data engineers. Clarifying this is necessary for understanding how language technologies can benefit male and female data engineers to the same extent.

3. Formal ontologies

What is needed, then, is a framework that adapts fairness ideals to NLP context and allows for their critical examination in a systematic manner. FOs provide such a framework, because they force the modeler to make conceptual commitments explicit and concrete. They typically represent concepts as classes whose meanings are specified through logical axioms expressing necessary (and sometimes sufficient) conditions for class membership, together with their relations to other classes and additional logical constraints [11, 18, 19]. The FO's axioms thus formalize conceptual commitments, and support logical inferences about class membership and relations between classes. As such, constructing a FO forces the modeler to make choices regarding how concepts are defined, distinguished, and related to other concepts, and these choices will remain explicitly visible in the axioms of the FO.

Since FOs explicitly encode conceptual commitments, these commitments can be critically examined. This makes FOs particularly suitable for domains in which conceptual and normative considerations are intertwined, such as NLP bias mitigation. Choices that are often left implicit in existing methods, such as which gender distinctions should be preserved and which should be removed, can instead be represented explicitly in the ontology. For example, a FO of occupations requires an explicit decision about whether "male data engineer" and "female data engineer" are to be represented as distinct concepts or as instances of the same concept with different values for the property of sex. If they are represented as distinct concepts, the FO must also specify whether they differ in defining properties other than sex.

The value of FOs lies *not* in providing answers to normative choices (they don't), but rather in making such choices explicit. If these choices are systematically decided according to some normative stance such as a certain conception of fairness, then the FO provides an operationalization of that fairness ideal. As an added value, FOs force choices that abstract normative stances such as conceptions of fairness

leave underdetermined. For example, the normative stance that gender should not be inferable from occupation does not by itself determine how lexically gendered occupational terms (e.g., "policeman", "midwife") or occupations subject to legal or institutional gender restrictions (e.g., "Catholic priest") ought to be represented. Constructing a FO exposes such unresolved choices and requires the modeler to confront them.

As such, FOs provide a framework for explicitly representing what counts as legitimate semantic content according to the normative conception they embody. Whereas bias mitigation has largely focused on identifying and eliminating harmful or undesirable statistical associations, FOs make it possible to formulate a positive account of legitimate semantic content by explicitly representing which conceptual distinctions, properties, and relations ought to be reflected in language models.

The proposal is thus to use FOs to make the semantic assumptions underlying NLP bias mitigation explicit. More specifically, it is to use FOs as representations of legitimate semantic content under a given normative stance, making its consequences, assumptions, and possibly inconsistencies transparent, thereby allowing these semantic assumptions to be critically assessed. Following such evaluation, the FO can serve as the semantic baseline against which candidate NLP debiasing interventions can be evaluated.

4. FOs as an interdisciplinary methodology

Another important aspect of FOs is their interpretability. As explicit representations of concepts and relations structuring a domain, FOs are accessible not only to ontology engineers, but also to non-technical audiences such as humanities researchers.[2] This is important, because NLP researchers have repeatedly called for collaborations with humanities researchers (e.g., [5, 1, 8, 9, 22]). FOs provide a shared representational framework which facilitates collaboration across disciplines, enabling different forms of expertise to contribute to the design and evaluation of NLP systems. In particular, they enable humanists and NLP developers to articulate and compare their views of legitimate semantic content, fostering discussion of the assumptions and norms involved in NLP bias mitigation. Such an interdisciplinary role is already well established in domains such as biomedicine and cultural heritage, where FOs are routinely used to integrate knowledge from multiple disciplines through providing a shared terminology and conceptual framework [23, 24].

One aspect of NLP bias mitigation where interdisciplinary collaboration is particularly needed is the representation of gender. With a handful of exceptions (e.g., [22, 25, 26]), NLP bias mitigation generally assumes the "folk model" of gender, according to which gender is binary and determined by sex, and is hetero-normative [9, 27]. Besides committing the representational harm of ignoring non-binary gender identities and same-sex relationships, this also leads to problematic methodological assumptions. For example, reliance on the folk-model motivates practices such as swapping out pairs of gendered terms, such as "actor" for "actress", to assess differences across outputs [28]. The semantic assumption underlying such methods is that these pairs differ only in gender, however, this assumption fails: the female term may carry pejorative connotations absent in the male term (e.g., "bachelor", "spinster"). These terms are not, in fact, semantic equivalents differing only in gender (cf. [9, 27]).

Another gap at the heart of NLP gender bias research is the lack of an explicit conceptualization and shared vocabulary for gender. A recent survey of nearly 200 articles on gender bias in NLP showed that the majority of these articles do not make their theorization of gender explicit, and that many conflate sex characteristics, social gender, and grammatical gender [9]. For example, several papers use the terms "male" and "female" to describe grammatical or lexical gender, whereas linguistics traditionally distinguishes "masculine", "feminine" and "neuter", or numbers these classes [9]. Such terminological inconsistencies reflect deeper conceptual ambiguities, and make it difficult to relate NLP research to

[2] For non-technical audiences unfamiliar with the logical formalisms underlying FOs, accessibility may be enhanced by approximate representations in graphical languages (e.g. OntoUML [20]) or natural language annotations accompanying the formalization in ontology editors such as Protégé [21]. Moreover, the transparency of FOs does not come primarily from their logical syntax, but rather from established methodology of ontology development, which includes documenting definitions, discussing alternative conceptualizations, and validating the ontology with domain experts.

established work in other relevant disciplines such as linguistics, gender studies, and feminist philosophy. In addition, without an explicit conceptualization of gender, it remains unclear to whom gender bias in NLP systems might be harmful and in what way (cf. [8, 9]). Thus, an explicit conceptualization of gender is required to evaluate which aspects of gender language technologies should represent, and thus to decide which aspects should be eliminated by NLP bias mitigation.

In sum, FOs provide a valuable framework for fostering interdisciplinary collaborations. In particular, they can serve as a bridge between the humanities and NLP, thereby facilitating the much-needed collaboration between researchers in NLP and the humanities on bias mitigation in language technologies. Because they explicitly represent concepts and their relations, they are accessible across disciplines and can provide a common vocabulary and shared conceptual framework, including an explicit conceptualization of gender.[3] As such, FOs may help better align NLP bias research with work in other disciplines, such as (socio)linguistics, gender studies, and feminist philosophy, fostering a more conceptually grounded and interdisciplinary approach to NLP bias mitigation.

5. Conclusion and future work

NLP bias mitigation approaches need to distinguish regularities that reflect legitimate semantic content in language models from bias. How to do so is a complex conceptual and normative problem, not merely a technical one. This paper proposed to use FOs to make explicit the normative conceptions of legitimate semantic content that implicitly underlie NLP bias mitigation. FOs provide a methodology that enables explicit semantic theorizing, facilitates interdisciplinary collaboration, and supports the systematic evaluation of bias mitigation approaches. In doing so, they offer not merely a technical tool, but an interdisciplinary framework for integrating humanities expertise with NLP bias research and practice.

In future work, I aim to put this proposal into practice. That is, I aim to develop a FO of a social domain such as occupations that formalizes a normative stance on gender fairness, such as treating like cases alike. I aim to test how useful such a FO is for evaluating existing bias mitigation approaches, whether it can be integrated into existing NLP bias mitigation pipelines, and whether it is successful in facilitating interdisciplinary collaboration. In doing so, I aim to explore shifting the focus from identifying what should be removed from a language model to articulating what should count as legitimate semantic content.

Acknowledgments

The author thanks Viola Schiaffonati and Roberta Ferrario for valuable discussions that helped shape the ideas presented in this paper and for comments on an earlier draft that improved the paper.

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT (OpenAI, GPT 5.5) to improve writing style. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] S. Barocas, K. Crawford, A. Shapiro, H. Wallach, The problem with bias: Allocative versus representational harms in machine learning, in: 9th Annual conference of the special interest group for computing, information and society, volume 1, New York, NY, 2017.

[3] In fact, FOs are already used for the representation of gender in domains such as cultural heritage [29].

[2] P. Nemani, Y. D. Joel, P. Vijay, F. F. Liza, Gender bias in transformers: A comprehensive review of detection and mitigation strategies, Natural Language Processing Journal 6 (2024) 100047.
[3] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, N. K. Ahmed, Bias and fairness in large language models: A survey, Computational Linguistics 50 (2024) 1097–1179.
[4] A. T. Cignarella, A. Giachanou, E. Lefever, A survey on stereotype detection in natural language processing, ACM Computing Surveys 58 (2025) 1–33.
[5] M. Bartl, A. Mandal, S. Leavy, S. Little, Gender bias in natural language processing and computer vision: A comparative survey, ACM Computing Surveys 57 (2025) 1–36.
[6] D. Hellman, Algorithmic Fairness, in: E. N. Zalta, U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy, Fall 2025 ed., Metaphysics Research Lab, Stanford University, 2025.
[7] D. Hovy, S. L. Spruit, The social impact of natural language processing, in: K. Erk, N. A. Smith (Eds.), Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Berlin, Germany, 2016.
[8] S. L. Blodgett, S. Barocas, H. Daumé Iii, H. Wallach, Language (technology) is power: A critical survey of "bias" in nlp, arXiv preprint arXiv:2005.14050 (2020).
[9] H. Devinney, J. Björklund, H. Björklund, Theories of "gender" in nlp bias research, in: Proceedings of the 2022 ACM conference on fairness, accountability, and transparency, 2022, pp. 2083–2102.
[10] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, Advances in neural information processing systems 29 (2016).
[11] N. Guarino, D. Oberle, S. Staab, What is an ontology?, in: Handbook on ontologies, Springer, 2009, pp. 1–17.
[12] D. Plunkett, T. Sundell, Disagreement and the semantics of normative and evaluative terms (2013).
[13] H. Cappelen, Fixing language: An essay on conceptual engineering, Oxford University Press, 2018.
[14] M. Kaneko, D. Bollegala, Gender-preserving debiasing for pre-trained word embeddings, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1641–1650.
[15] A. P. Association, et al., Publication manual of the american psychological association,(2020), American Psychological Association 428 (2019).
[16] A. Piergentili, B. Savoldi, D. Fucci, M. Negri, L. Bentivogli, Hi guys or hi folks? benchmarking gender-neutral machine translation with the gente corpus, arXiv preprint arXiv:2310.05294 (2023).
[17] A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, Science 356 (2017) 183–186.
[18] C. Masolo, S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, Wonderweb deliverable d18 ontology library (final), ICT project 33052 (2003).
[19] S. Borgo, C. Masolo, Foundational choices in dolce, in: Handbook on ontologies, Springer, 2009, pp. 361–381.
[20] J. Guerson, T. P. Sales, G. Guizzardi, J. P. A. Almeida, Ontouml lightweight editor: A model-based environment to build, evaluate and implement reference ontologies, in: 2015 IEEE 19th International Enterprise Distributed Object Computing Workshop, 2015, pp. 144–147. doi:10.1109/EDOCW.2015.17.
[21] M. A. Musen, The protégé project: a look back and a look forward, AI matters 1 (2015) 4–12.
[22] Y. T. Cao, H. Daumé III, Toward gender-inclusive coreference resolution, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020.
[23] S. Schulz, L. Jansen, Formal ontologies in biomedical knowledge representation, Yearbook of medical informatics 22 (2013) 132–146.
[24] M. Doerr, Ontologies for cultural heritage, in: Handbook on ontologies, Springer, 2009, pp. 463–486.
[25] B. Savoldi, M. Gaido, L. Bentivogli, M. Negri, M. Turchi, Gender bias in machine translation, Transactions of the Association for Computational Linguistics 9 (2021) 845–874.

[26] T. Sun, K. Webster, A. Shah, W. Y. Wang, M. Johnson, They, them, theirs: Rewriting with gender-neutral english, arXiv preprint arXiv:2102.06788 (2021).
[27] F. Calado, Some myths about bias: A queer studies reading of gender bias in nlp, in: Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP), 2025, pp. 338–346.
[28] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, K.-W. Chang, Gender bias in coreference resolution: Evaluation and debiasing methods, arXiv preprint arXiv:1804.06876 (2018).
[29] S. Anastasi, A. De Domenico, M. N. Asmundo, Egith-engendering data. steps towards an ontology for the representation of gender in cultural heritage., in: AI4CH@ AI* IA, 2022, pp. 13–25.