Challenging “Subgroup Fairness”: Towards Intersectional Algorithmic Fairness based on Personas

Marie Christin Decker[1,*] Laila Wegner[1] and Carmen Leicht-Scholten[1]

[1]RWTH Aachen University, Templergraben 55, 52062 Aachen, Germany

Abstract

To react to social injustices and harmful stereotypes caused by algorithms, fairness metrics evaluate how well an algorithm performs for individuals or sets of groups. However, judging based on so-called protected attributes squeezes people’s characteristics into fixed structures and ignores the socially constructed nature of human identities. A more founded way to conceptualize the complexity of human identities demands to engage in interdisciplinary theories from philosophy, sociology, and gender studies on intersectionality and social identity. In this contribution, we sketch how current approaches to algorithmic fairness fall short in considering human identities and intersectional realities. We propose an approach based on personas for a more holistic view of humans and intersectionality in algorithmic fairness considerations.

Keywords

Machine Learning, Algorithmic Fairness, Intersectionality, Identity, Personas

1. Introduction

Given that algorithmic decision-making (ADM) can reinforce societal stereotypes and systematic disadvantages, fairness metrics evaluate how well an algorithm performs for individuals (individual fairness) or sets of groups (group fairness) (for discussion, see e.g., [1]). While computational fairness approaches are valuable in a diagnostic function and draw attention to long-standing social injustices [2], critical scholars emphasized several weaknesses of algorithmic fairness [3]. One point of criticism is that the common algorithmic fairness metrics evaluate discrimination with respect to predefined protected attributes as they are listed for example in the General Equal Treatment Act in Germany (AGG) [4], the non-discrimination article by the European Commission [5] or Title VII of the Civil Rights Act of 1964 in the U.S. [6]. These acts list explicit attributes such as Race, Color, Ethnicity, Gender, Sex, Religion, (Dis)Ability, or Age based on which fairness metrics evaluate whether differential treatment is present in an algorithm. This is criticized because people’s characteristics are squeezed into and judged based on fixed attributes, ignoring the socially constructed nature of human identities [7]2. A more founded way to conceptualize the complexity of human identities demands to engage in interdisciplinary theories from philosophy, sociology, and gender studies on intersectionality and social identity.

In this contribution, we sketch how current approaches to algorithmic fairness fall short in considering human identities and intersectional realities. We propose an approach based on personas

EWAF'24: European Workshop on Algorithmic Fairness, July 01–03, 2024, Mainz, Germany
* Corresponding author.
marie.decker@gdi.rwth-aachen.de (M. Decker); laila.wegner@rwth-aachen.de (L. Wegner); carmen.leicht@gdi.rwth-aachen.de (C. Leicht-Scholten)
0000-0002-9138-231X (M. Decker); 0000-0001-9983-8361 (L. Wegner); 0000-0003-2451-6629 (C. Leicht-Scholten)

[2] In addition to substantive criticism, there is also criticism of the term ‘protected attributes’ itself because it evokes the impression that certain groups of people need (external) protection [8].

from human-computer interaction for a more holistic view of humans and intersectionality in algorithmic fairness considerations that we plan to explore further in future research.

2. Forgotten Identities in Algorithmic Fairness

Concerns regarding algorithmic fairness are mostly grounded in the observation that algorithms discriminate against certain attributes. However, Heinrichs [9] highlights that claims of wrongful discrimination must be supported by ethical reasoning about what exactly makes it wrong. Evaluating algorithmic discrimination and determining what makes an algorithm fair therefore also requires an agreement on the underlying normative assumptions. To base this decision only on protected attributes (as is often done in technical contexts) implies that a straightforward distinction between unacceptable (= those based on protected attributes) and acceptable inequalities (= those not based on protected attributes) is possible [10]. For example, discrimination based on math skills for a math job seems intuitively legitimate. At the same time, most people would disagree that a feature such as gender should influence the opportunity to work in a math job. Depending on what counts as morally wrong, the list of protected attributes could be endless. While the law is certainly a first orientation, critical scholars question whether it is fully suitable and sufficient for the context of algorithmic discrimination [11].

The assumption that protected attributes can distinguish between wrongful and acceptable discrimination is even more challenged by the socially constructed nature of categories [12]: concepts such as gender or race are not solely descriptions of ones belonging to a fixed class, nor can they be seen as natural or objective, but instead contain deeply inherited social values [13–16]. By (computationally) reducing this ascribed social meaning to calculable categories, fluid concepts are conceptualized as natural and objective facts [15, 17] – particularly neglecting the lived nature of experiences that come along with belonging to more than one of these categories. Ignoring these diverse experiences can reinforce stereotypes and stigmatization because it evokes the impression that some group members (e.g., females) need protection and help in contrast to those more privileged (e.g., males). Hoffmann highlights that a focus on discrimination and disadvantage additionally shifts away discourses from the acquisition of privileges [11].

These challenges point already to the limited representation of identity in the context of using protected attributes for algorithmic fairness. As highlighted by Crawford [18, p. 147], the categorization of humans into fixed categories predefines the scope of the very dynamic and relational nature of human identity and "restricts the range of how people are understood and can represent themselves, and it narrows the horizon of recognizable identities.". The consequence is a loss of personal identification [14]: the observable characteristic 'skin color' does probably not match the individual self-identification of a group which may be based on cultural traditions and the social context [19]. A social category becomes a collective identity only if it is personally acknowledged – a challenge when identities are summarized in fixed protected attributes. This is further underlined by the Social Identity Theory coined by Tajfel [20] stressing the dynamic nature of human identity. Briefly, the theory describes that people's selves are made up of different identities. Identities are significantly shaped by membership to one or multiple social group(s) – including the values and norms that come along with these memberships. Only some of these identities relate to the protected attributes defined by law [19, 21, 22], other groups include for example sports clubs or professions [23].

That identities are complex and need to be considered in their interplay [24, 25] is particularly highlighted in the notion of intersectionality coined by Crenshaw [26, 27]. In a nutshell, intersectionality describes the overlap of systems of oppression, making people experience discrimination on more than one level. The societal separation between privilege and disadvantage follows individual attributes such as race, gender, or social class [28] which reflect patterns of status and power and are turned into oppression (e.g., sexism, racism). Consequently, intersectional discrimination cannot be considered the sum of multiple single discriminatory experiences [26, 27]. Intersectionality further highlights that those privileged determine the distribution of power. Privileges refer to unearned advantages that people are entitled to because they belong to a certain social group or have certain dimensions of their identity [29]. By exercising power, the privileged group determines social norms, setting the normative ideal and creating a picture of the 'others' [29].

There have been approaches to account for intersectionality in the algorithmic fairness literature, most of which compare parity notions of subgroups (e.g., Black women) [30]. Questions deal for example with the correct number and type of subgroups. While these approaches are important to make intersectional biases visible, an additional engagement with power dynamics inherent in social categories is necessary to truly account for intersectionality in a sociological sense [11]. Creating equalized subgroups may not emphasize that discriminations based on several attributes mutually intensify each other and are not equal to the sum of individual experiences [13, 16]. Therefore, Kong [30, p. 492] demands a shift of "the focus of fairness research from intersections of protected attributes to intersections of structural oppression".

While intersectionality is not directly analyzed with respect to the interplay of social identity, critical scholars highlight the shortcomings in acknowledging the fluid and dynamic nature of human identity in algorithmic fairness literature generally. For example, Lu et al. [31] argue that static, non-relational categories of humans immobilize concepts of human identity and codify existing norms. Motivated by a critical lens on protected attributes, Belitz et al. [32] develop context-specific categories based on sociological approaches to identity. To do so, they incorporate self-identifications into new classifications to bring the algorithmic representation closer to the self-concept of identity. Considering self-identifications is an important step to include different social identities in algorithmic fairness research. However, they derive categories from self-categorization which risks becoming analyzed in isolation and thus may reproduce some of the shortcomings of protected attributes. As we sketch in the following, our approach differs because we aim to consider a connected abstraction of identity instead of categories by using personas.

3. Personas to Conceptualize Identity in Algorithmic Fairness

Building on the previous work on human identity in algorithmic fairness and informed by a critical intersectional framework, we propose to target the coexistence of multiple social identities and its consequences to conceptualize the whole complexity of algorithmic fairness. Therefore, we draw on social identity theory and propose context-specific, participatively developed personas as an approach to enhance intersectional algorithmic fairness. While personas were considered in [33], their contribution focuses on AI in general and its implication for explainable AI. However, to the best of our knowledge, the approach to include personas in algorithmic fairness research is novel.

Personas represent symbolic and fictive people vignettes that are developed to understand user-specific requirements and interests, for example in human-computer interaction [34, 35]. To provide an example, Marsden and Pröbster [34, p. 10] propose the persona 'Lea' for the design context of an e-learning platform for women in tech: "We built a persona called "Lea" with a high career orientation (I5), with a supportive husband who works parttime and two children. This persona is ambitious, very organized, strategic, and interested in networking. She also networks in a very strategic way, reaching out to people who can help her. Gender stereotypes annoy her and she cannot relate to any of them". Personas are used as representants for larger groups, combining individual attributes and social categories. Unlike protected attributes, they target people's goals, interests, and behavior for the relevant context [36]. Therefore, personas should not be simplified by merging them into categories like gender, race, or age, because this reproduces stereotypes and loses the diversity personas shall bring [34]. As a connection between user and designer, they incentivize designers to take on the perspective of different users in the design process [36–39] and to directly adapt the design to the needs of the users [40].

In our research project, we wish to explore how personas can serve as an approach to integrate the co-existence of multiple identities and thus, to represent the variety of human identity, into algorithmic fairness research. We imagine the use of personas in all stages of the Machine Learning cycle: For reflection in the design phase, for explicit consideration in the building phase as well as for evaluation in the testing and monitoring phase. By using personas in these contexts, we anticipate the following benefits: First, the use of personas is aimed to partly serve as an answer to the weaknesses of protected attributes: Although personas are still an abstraction and simplification of the reality, they do not replicate the separated analysis of individual attributes but represent a combination of different identities. In this way, personas can depict intersectional experiences contrary to the traditional abstractions in algorithmic fairness. By designing personas in a fluid fashion, the binary and mutually exclusive narrative of protected attributes is aimed to be challenged and individual experiences are included in the abstract representation of humans. Consequently, intersectional personas represent different realities of life within legally protected groups and include more social identities in their different contexts. Thus, personas are a step towards meaningfully combining individual and group-based fairness metrics and can help to make these intra-group differences more visible while still being generalizable. At best, a restricted focus on inter-group comparisons which neglects differences within groups [41] might be avoided.

However, personas do not only combine different protected attributes. Instead, they create a picture of imaginary real-life persons which “is meant to decrease our reliance on our own egocentric perspective when reasoning about other people’s thoughts, feelings, and other subjective experiences.” [34, p. 2]. This can be valuable to foster forward-looking responsibility in the context of algorithmic fairness. Forward-looking responsibility describes the engagement with future events and harm which shall be avoided in the first place. As discussed by Santoni de Sio and Mecacci [42, p. 1067], algorithms introduce an active responsibility gap which creates an obstacle to forward-looking responsibility. They stress that “engineers and other agents involved in the development and use of technology may not be (fully) aware of their respective moral and social obligations towards other agents”. Personas can add value by highlighting that algorithmic decision-making is not only a neutral data collection but directly influences the reality of human beings. In this way, the humans behind the data become visible and developers are reminded of the high stake and impact of their decisions in the respective context.

Finally, meaningful personas might be used as an approach to challenge algorithmic power hierarchies, which is one fundamental concern of intersectionality. To account for this, it is necessary to question how – and by whom – identity is conceptualized, and what the underlying norms are. Given the current power imbalances, a dominant group establishes societal norms around themselves by defining other groups as the inferior deviation [43]. For the development of algorithmic systems, this is especially relevant given current discussions around the formation of a coding elite [44]. This coding elite is predominated by a homogenous and privileged group (western-centralized [15, 45–47]), leading to the fact that their perspectives and experiences are built into the algorithms [15, 46, 47].

To avoid justice evaluations without considering those affected, it is necessary to approach these questions from a bottom-up perspective [48]. Personas might become a means to represent the variety of human experiences if they are developed in cooperation with those affected, for example, based on participatory approaches. Participatory personas ought to include marginalized perspectives so that they become a representative to give everyone a voice – at least in the sense that everyone in the given context can self-identify broadly with one of the developed personas. By valuing lived experiences, meaningful personas may become a means to reduce relational and epistemic injustices and acknowledge power hierarchies. This means also including personas that represent rather privileged identities, to overcome the single-axis focus on discrimination in algorithmic fairness research, as demanded by [11]. Thus, personas can help to question privileges instead of taking them for granted and only highlighting disadvantages. Engaging with privileges is an important part of embracing forward-looking responsibility as demanded above3.

Constructing meaningful personas is not without challenges. Personas need to simplify and abstract complex social realities without risking reproducing stereotypes or creating a false sense of understanding [34, 50]. Further, personas depend heavily on the context of use. For example, people can have different attitudes and experiences in their job and private life. This makes personas resource-intensive and costly to construct, especially when using qualitative or participatory approaches.

However, we are convinced that future research in algorithmic fairness should engage with overcoming the sketched oversimplified assumptions of human identity that create fixed norms and restrict humans in their diversity. Further, future research should engage with approaches to intersectionality that can account for power hierarchies without rebuilding assumptions of privileges into fairness evaluation, mirroring stigmatizing group hierarchies. In this short contribution, we proposed to use personas to account for these challenges and motivate our research proposal by the outlined advantages that come into play when limited protected attributes are replaced by meaningful personas. In future work, we plan on designing participative personas and test different scenarios on how to integrate them in the Machine Learning cycle.

Acknowledgments

This research was conducted within the RRI Hub at RWTH Aachen University. The RRI Hub is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of Germany's Excellence Strategy.

References

[1] R. Binns, On the Apparent Conflict between Individual and Group Fairness, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 514–524.
[2] R. Abebe, S. Barocas, J. Kleinberg, K. Levy, M. Raghavan, and D. G. Robinson, Roles for computing in social change, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona Spain, 2020, pp. 252–260.
[3] L. Weinberg, Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches, Journal of Artificial Intelligence Research, volume 74, pp. 75–109, 2022.
[4] Allgemeines Gleichbehandlungsgesetz: AGG, 2006. URL: https://www.antidiskriminierungsstelle.de/SharedDocs/downloads/DE/publikationen/AGG/agg_gleichbehandlungsgesetz.pdf?__blob=publicationFile
[5] European Union, Article 21 - Non-discrimination, in: Charter of Fundamental Rights of the European Union, European Union Agency for Fundamental Rights, Ed., 2012, p. 10.
[6] Title VII of the Civil Rights Act of 1964: Pub. L. 88-352 (Title VII). URL: https://www.eeoc.gov/statutes/title-vii-civil-rights-act-1964
[7] S. Barocas, M. Hardt, and A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities: MIT Press, 2023.
[8] L. M. Hampton, Black Feminist Musings on Algorithmic Oppression, in: Conference on Fairness, Accountability, and Transparency (FAccT '21), Virtual Event, Canada, 2021.
[9] B. Heinrichs, Discrimination in the age of artificial intelligence, AI & SOCIETY, 2021, doi: 10.1007/s00146-021-01192-2.
[10] M. S. A. Lee, L. Floridi, and J. Singh, Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics, AI Ethics, volume 1, no. 4, pp. 529–544, 2021, doi: 10.1007/s43681-021-00067-y.

[3] The role of privileges in responsibility was impressively underlined in the Social Connection Model by Iris Young [49], highlighting that social position increases the moral responsibility to address structural injustices.

[11] A. L. Hoffmann, Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse, Information, Communication & Society, volume 22, no. 7, pp. 900–915, 2019, doi: 10.1080/1369118X.2019.1573912.
[12] A. Hanna, E. Denton, A. Smart, and J. Smith-Loud, Towards a critical race methodology in algorithmic fairness, in: Conference on Fairness, Accountability and Transparency (FAT* '20), Association for Computing Machinery, Ed., New York, NY, 2020, pp. 501–512.
[13] T. Krupiy, A vulnerability analysis: Theorising the impact of artificial intelligence decision-making processes on individuals, society and human diversity from a social justice perspective, Computer Law & Security Review, volume 38, p. 105429, 2020, doi: 10.1016/j.clsr.2020.105429.
[14] M. Andrus and S. Villeneuve, Demographic-Reliant Algorithmic Fairness: Characterizing the Risks of Demographic Data Collection in the Pursuit of Fairness, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul Republic of Korea, 2022, pp. 1709–1721.
[15] S. Leavy, E. Siapera, and B. O'Sullivan, Ethical Data Curation for AI: An Approach based on Feminist Epistemology and Critical Theories of Race, in: AIES '21, Virtual Event, USA, 2021, pp. 695–703.
[16] A. Zimmermann and C. Lee-Stronach, Proceed with Caution, Canadian Journal of Philosophy, volume 52, no. 1, pp. 6–25, 2022, doi: 10.1017/can.2021.17.
[17] B. Green, Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness, Philosophy & Technology, volume 35, no. 4, 2022, doi: 10.1007/s13347-022-00584-6.
[18] K. Crawford, Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence: Yale University Press, 2021.
[19] R. D. Ashmore, K. Deaux, and T. McLaughlin-Volpe, An organizing framework for collective identity: articulation and significance of multidimensionality, Psychological bulletin, volume 130, no. 1, pp. 80–114, 2004, doi: 10.1037/0033-2909.130.1.80.
[20] H. Tajfel, Social categorization, social identity and social comparison, Differentiation between social groups: Studies in the social psychology of intergroup relations, pp. 61–76, 1978. URL: https://cir.nii.ac.jp/crid/1571980075816748032
[21] M. A. Hogg, D. Abrams, S. Otten, and S. Hinkle, The Social Identity Perspective, Small Group Research, volume 35, no. 3, pp. 246–276, 2004, doi: 10.1177/1046496404263424.
[22] M. A. Hogg and D. J. Terry, Social Identity and Self-Categorization Processes in Organizational Contexts, The Academy of Management Review, volume 25, no. 1, p. 121, 2000, doi: 10.2307/259266.
[23] D. Scheepers and N. Ellemers, Social identity theory, Social psychology in action: Evidence-based interventions from theory to practice, pp. 129–143, 2019, doi: https://doi.org/10.1007/978-3-030-13788-5_9
[24] A. van Dommelen, K. Schmid, M. Hewstone, K. Gonsalkorale, and M. Brewer, Construing multiple ingroups: Assessing social identity inclusiveness and structure in ethnic and religious minority group members, Euro J Social Psych, volume 45, no. 3, pp. 386–399, 2015, doi: 10.1002/ejsp.2095.
[25] E. M. Trauth, C. C. Cain, K. D. Joshi, L. Kvasny, and K. M. Booth, The Influence of Gender-Ethnic Intersectionality on Gender Stereotypes about IT Skills and Knowledge, SIGMIS Database, volume 47, no. 3, pp. 9–39, 2016, doi: 10.1145/2980783.2980785.
[26] K. Crenshaw, Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics, University of Chicago Legal Forum, volume 1989, no.8, pp. 139 – 167, 1989
[27] K. Crenshaw, Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color, Stanford Law Review, volume 43, no. 6, p. 1241, 1991, doi: 10.2307/1229039.
[28] E. Anderson, What is the Point of Equality?, Ethics, volume 109, no. 2, pp. 287–337, 1999, doi: 10.1086/233897.
[29] P. McIntosh, White privilege: Unpacking the invisible knapsack, in: Multictlturalism, A. M. Filor, Ed., 1992, pp. 30–36.
[30] Y. Kong, Are "Intersectionally Fair" AI Algorithms Really Fair to Women of Color? A Philosophical Analysis, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul Republic of Korea, 2022, pp. 485–494.

[31] C. Lu, J. Kay, and K. R. McKee, Subverting machines, fluctuating identities: Re-learning human categorization, in: FAccT '22, 2022, pp. 1005–1015.
[32] C. Belitz, J. Ocumpaugh, S. Ritter, R. S. Baker, S. E. Fancsali, and N. Bosch, Constructing categories: Moving beyond protected classes in algorithmic fairness, Journal of the Association for Information Science and Technology, volume 74, no. 6, pp. 663–668, 2022, doi: 10.1002/asi.24643.
[33] A. Holzinger, M. Kargl, B. Kipperer, P. Regitnig, M. Plass, and H. Muller, Personas for Artificial Intelligence (AI) an Open Source Toolbox, IEEE Access, volume 10, pp. 23732–23747, 2022, doi: 10.1109/ACCESS.2022.3154776.
[34] N. Marsden and M. Pröbster, Personas and Identity, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow Scotland Uk, 2019, pp. 1–14.
[35] J. S. Pruitt and T. Adlin, Eds., The persona lifecycle: Keeping people in mind throughout product design. Amsterdam, Boston: Elsevier, 2006.
[36] J. Grudin, WHY PERSONAS WORK: THE PSYCHOLOGICAL EVIDENCE, in: The Morgan Kaufmann series in interactive technologies, The persona lifecycle: Keeping people in mind throughout product design, J. S. Pruitt and T. Adlin, Eds., Amsterdam, Boston: Elsevier, 2006, pp. 642–663.
[37] L. Nielsen, Personas: User focused design. London, Heidelberg: Springer, 2013.
[38] J. Pruitt and J. Grudin, Personas, in: Proceedings of the 2003 conference on Designing for user experiences, San Francisco California, 2003, pp. 1–15.
[39] A. Cooper, The inmates are running the asylum: Why high-tech products drive us crazy and how to restore the sanity, 6th ed. Indianapolis, Ind.: Sams, 2006.
[40] M. C. Jones, I. R. Floyd, and M. B. Twidale, Teaching Design with Personas, Interaction Design and Architecture(s), 3-4, pp. 75–82, 2007.
[41] V. M. May, Pursuing intersectionality, unsettling dominant imaginaries. New York: Routledge, 2015. URL: https://www.taylorfrancis.com/books/9781136497551
[42] F. Santoni de Sio and G. Mecacci, Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them, Philosophy & Technology, volume 34, no. 4, pp. 1057–1084, 2021, doi: 10.1007/s13347-021-00450-x.
[43] I. M. Young, Ed., Justice and the politics of difference. Princeton: Princeton University Press, 1990.
[44] J. Burrell and M. Fourcade, The Society of Algorithms, Annual Review of Sociology, volume 47, no. 1, pp. 213–237, 2021, doi: 10.1146/annurev-soc-090820-020800.
[45] A. Birhane, Algorithmic injustice: a relational ethics approach, Patterns (New York, N.Y.), volume 2, no. 2, 2021, doi: 10.1016/j.patter.2021.100205.
[46] L. T.-L. Huang, H.-Y. Chen, Y.-T. Lin, T.-R. Huang, and T.-W. Hun, Ameliorating Algorithmic Bias, or Why Explainable AI Needs Feminist Philosophy, Feminist Philosophy Quarterly, volume 8, ¾, 2022.
[47] L. M. Rafanelli, Justice, injustice, and artificial intelligence: Lessons from political theory and philosophy, Big Data & Society, volume 9, no. 1,2022.
[48] P.-H. Wong, Democratizing Algorithmic Fairness, Philosophy & Technology, volume 33, no. 2, pp. 225–244, 2020, doi: 10.1007/s13347-019-00355-w.
[49] I. M. Young, Responsibility and global justice: A social connection model, Social philosophy and policy, volume 23, no. 1, pp. 102–130, 2006.
[50] N. Marsden and M. Haag, Stereotypes and Politics, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose California USA, 2016, pp. 4017–4031.