Invisible Inequalities – Intersectional Fairness in Educational AI

MARIE MIRSCH, RWTH Aachen University, Germany
JONAS STRUBE, RWTH Aachen University, Germany
CARMEN LEICHT-SCHOLTEN, RWTH Aachen University, Germany

Drawing on feminist theories of Intersectionality, this paper explores how single-axis approaches to fairness assessments obscure the experiences of individuals facing intersecting forms of discrimination. Three case studies in educational AI illustrate how individuals' social embeddedness shapes their educational trajectories and why fairness metrics often fail to account for these complexities. The paper argues that addressing invisible inequalities requires a shift from purely technical solutions to context-sensitive fairness evaluations that center on the lived experiences of marginalized people.

Keywords: Fairness, Intersectionality, Machine Learning, Artificial Intelligence, Educational AI

Reference Format:
Marie Mirsch, Jonas Strube, and Carmen Leicht-Scholten. 2025. Invisible Inequalities – Intersectional Fairness in Educational AI. In *Proceedings of the Fourth European Workshop on Algorithmic Fairness (EWAF'25)*. Proceedings of Machine Learning Research, 8 pages.

1 Introduction

Numerous technical methods have been proposed to mitigate bias and thus entrench algorithmic fairness in Machine Learning (ML) systems. These include statistical approaches that assess fairness through outcome parity across demographic groups [4], individual-based approaches that evaluate fairness by comparing outcomes for similar individuals [19], and causal approaches that aim to minimize the influence of demographic attributes on outcomes [30]. Although all of these are important, they face substantial criticism for being resource-intensive, lacking well-defined similarity measures, depending on structural choices, or overlooking procedural and relational perspectives [18].

In this paper, we highlight one particular point of criticism: the widespread practice of conducting fairness evaluations for only one sensitive attribute at a time [15, 17, 26, 31, 38]. This practice echoes what Crenshaw [1989] criticized in the US legal system's treatment of Black women's discrimination claims: The legal framework's single-axis approach to discrimination failed to recognize how multiple forms of marginalization can intersect, leaving those who experience discrimination at the intersection of race and gender (i.e., Black women) without adequate protection. More than 35 years later, despite the rapid rise of fairness considerations in technical systems and some efforts to account for multiple subgroups at a time [22], addressing fairness continues to obscure the experiences of individuals facing various, intersecting forms of discrimination. The feminist notion of *Intersectionality* describes how individuals' intersecting experiences are shaped by their embeddedness in specific social, cultural, and political contexts [13].

Authors' Contact Information: Marie Mirsch, RWTH Aachen University, Aachen, Germany, marie.mirsch@gdi.rwth-aachen.de; Jonas Strube, RWTH Aachen University, Aachen, Germany, jonas.strube@gdi.rwth-aachen.de; Carmen Leicht-Scholten, RWTH Aachen University, Aachen, Germany, carmen.leicht@gdi.rwth-aachen.de.

EWAF'25, June 30–July 02, 2025, Eindhoven, NL

In the domain of *educational AI,* these considerations become particularly urgent. At the latest, since the introduction of generative AI systems and the widespread adoption of learning analytics, educational AI systems alter educational landscapes by being used for high-stakes decision-making, such as predicting students' learning success [40] or providing adaptive diagnosis and feedback to students' learning progress [25] – with profound impacts on individuals' life chances. At the same time, educational settings are deeply embedded in broader social inequalities, where intersections of race, gender, disability, and socioeconomic background profoundly shape students' academic trajectories [33, 41]. In this field, single-axis approaches to fairness evaluations are still common [3].

To account for intersectional fairness requires profound discussion of feminist discourses on Intersectionality. Since understanding is a necessary prerequisite for action [34], drawing on Collins and Bilge [2016]'s notion of the six core ideas of Intersectionality, we discuss three case studies in the field of educational AI as a context-specific illustration of the relevance of structural considerations and intersectional reflections.

2 Intersectionality

Being a widely discussed notion in feminist theories, Intersectionality loosely describes the "critical insight that race, class, gender, sexuality, ethnicity, nation, ability, and age operate not as unitary, mutually exclusive entities, but rather as reciprocally constructing phenomena that in turn shape complex social inequalities" [12]. To make this definition more tangible, Collins and Bilge [2016] identify six core ideas that "appear and reappear" throughout Intersectional perspectives: *Complexity* recognizes that multiple intersecting identities – in the context of discrimination summarized in sensitive attributes [1] – shape individuals' identities. For example, a female student from an underrepresented racial group may face distinct challenges in educational AI systems compared to male students or peers from racially majority groups. *Relationality* highlights how these identity traits influence one another. The significance of gender, for instance, may shift when considered alongside age or socioeconomic status. *Social Context* stresses that the relevance of identity traits depends on situational environments. Language background may matter in language assessments, while gender may play a significant role in male-dominated STEM classrooms. *Social Inequalities* address how disparities emerge from different treatments based on intersecting identity traits. Educational AI systems can, when badly used, reinforce social inequalities – such as predictive models assigning lower scores to low-income students. *Power* is a central tenet of Intersectionality, drawing attention to the broader social hierarchies that privilege certain groups over others. Educational AI systems operate within these power dynamics, and fairness evaluations must critically examine how algorithmic decisions reinforce or challenge these power imbalances. Finally, *Social Justice* represents the normative aim of intersectional analyses. Beyond technical interventions to mitigate algorithmic bias, intersectional approaches advocate for addressing the underlying social and institutional power structures that produce inequality.

3 Intersectionality in Educational AI: Case Studies

We distinguish three dimensions of educational AI fairness. *Predictive AI systems* use historical data to forecast outcomes. Ensuring intersectional fairness in these systems requires mitigating disparities between multiply marginalized and privileged groups. Beyond predictive systems, concerns about fairness also arise in *generative AI systems,* where representational fairness pertains to avoiding harmful stereotypes or the underrepresentation of certain multiply marginalized groups. Lastly, *access and usage* fairness addresses the digital divide – inequalities in access to technology that impact opportunities to benefit from AI systems [9].

3.1 Predictive AI: Grade Prediction Models to Support Students at Risk

Grade prediction models are designed to support at-risk students by leveraging historical data, such as prior academic performance and demographic information. These models may notify selected students about available support resources [27] or provide real-time estimates of passing probabilities [2]. Jiang and Pardos [2021] demonstrated that incorporating race as an explicit factor in a particular grade prediction tool improved its overall accuracy. However, they also found that different racial groups experienced varying error rates, with underrepresented minorities generally receiving lower-quality predictions compared to groups with higher representation.

This discrepancy raises concerns, as minority students tend to be disadvantaged in educational settings – a *social inequality* that can be pertained or reinforced through grade prediction models. Research, for example, indicates that Black and Hispanic students perform lower in key subjects compared to their White and Asian peers [20, 42]. Beyond race alone, multiple identity traits, such as socioeconomic status or gender, intersect to influence grade predictions [28]. For instance, minority students are disproportionately likely to come from lower socioeconomic backgrounds, and economic disadvantages correlate with lower academic performance and higher dropout rates [36, 42]. A reductionist approach that considers these factors in isolation risks obscuring deeper patterns of disadvantage, whereas an intersectional perspective would emphasize the *relational* and *complex* nature of these attributes. Additionally, minority students are more likely to drop out or be placed in lower-status educational tracks, a trend observed across different countries and educational systems [6]. The accompanying stereotypes regarding the academic potential of minority groups can create barriers, contributing to anxiety, low self-esteem, and increased cognitive load, all of which negatively impact learning outcomes [24]. Misclassifying students from marginalized backgrounds as high-risk could discourage them, reinforcing existing educational inequalities through self-fulfilling prophecies. An intersectional perspective reveals how stereotyping is intertwined with *power relations* [39]. Finally, a *context-aware* model would discuss social conditions that generate these disparities instead of reducing fairness evaluation to attributes for which data is available.

3.2 Generative AI: GPT Detectors to Prevent Academic Dishonesty

Since generative AI systems introduce new complexities into educational contexts [37], GPT detectors have emerged as a (thoroughly controversial) tool for detecting AI-generated texts to prevent academic dishonesty. However, these tools have demonstrated a tendency toward misclassification, particularly for non-native (English) speakers [29, 32]. A study by Liang et al. [2023] revealed disparities in classification accuracy, with non-native English writing samples being misclassified as AI-generated 61.3% of the time, compared to 5.1% for native English essays. The researchers hypothesized that non-native speakers tend to use a more predictable linguistic structure with lower variability and limited word choice, making their writing more susceptible to AI misclassification.

This misclassification compounds existing systemic challenges and *structural inequalities* that non-native speakers face in education, such as difficulties in understanding classroom lectures and academic material [21, 43] or issues related to social integration [7]. As a result, affected students may face unfair accusations of academic dishonesty, lower grades, and increased scrutiny, further impacting their academic performance and social belonging. The intertwinement of linguistic proficiency with other (partly *related*) factors like socioeconomic status, immigrant background, or special education needs adds *complexity* to fairness evaluations [8]. Furthermore, classifying individuals as ‘native’ or ‘non-native’ is itself an exercise of *power* that shapes how individuals are perceived and treated [35]. For instance, the concept of a ‘native language’ is a Western construct that may not align with the lived experiences of individuals from diverse linguistic and cultural backgrounds [23], and native speakers who do not conform to Western linguistic norms, such as speakers of non-standard dialects, may face educational barriers [5]. Racial and ethnic biases may complicate linguistic categorization; in South Korea, for instance, teachers of color, regardless of their linguistic proficiency, may be perceived as non-native English speakers [11]. The importance of *context* in fairness evaluations by GPT detectors becomes evident in a study by Jiang et al. [2024] who did not report significant differences in GPT detector performance between native and non-native speakers. The researchers attributed this to their approach of training their detectors exclusively on the kind of essays they should evaluate rather than on a broad dataset. They argued that task-specific, purpose-built classifiers may enhance both accuracy and fairness compared to general-purpose GPT detectors.

3.3 Access & Usage: Digital Divide Predicting Computational Thinking Skills

Since unequal access to AI technologies can lead to significant gaps in usage, the third dimension, access and usage of AI in education, pertains to the availability of educational AI systems and the ability to engage with them effectively.

For example, white men are the most likely to have access to technological devices, whereas Black and Hispanic females are the least likely [44]. Further, individuals who do not engage with digital technology of any kind are more likely to be older and socially deprived [10]. These findings illustrate the *relational* and *complex* effects that shape access to digital technology. Further, *social inequalities* due to socioeconomic background, education level, and demographic attributes determine who benefits from technological advancements. For instance, Celik [2023] demonstrated a positive correlation between access to digital technology and AI literacy [9]. Consequently, disparities in access are likely to contribute to inequalities in AI literacy. This assumption is supported by studies identifying discrepancies in AI knowledge based on family’s socioeconomic status [45] and educational level [22]. Males and younger generations tend to be more knowledgeable about AI than females and older individuals [22]. An intersectional approach respects the fact that socioeconomic factors shape access to and engagement with AI systems.

4 Conclusion

Since current conceptions of algorithmic fairness primarily address discrimination through a single-axis lens, we discussed three case studies in the field of educational AI as a context-specific illustration of the relevance of structural considerations and intersectional reflections. Ultimately, an intersectional approach challenges the idea that fairness can be achieved solely through technical adjustments, instead advocating for a socially conscious approach that situates algorithmic fairness within broader structural and relational contexts. However, defining and operationalizing intersectional algorithmic fairness remains an ongoing challenge [16].

Acknowledgements

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] Antidiskriminierungsstelle des Bundes. 2006. *Allgemeines Gleichbehandlungsgesetz*: *AGG*.

[2] Kimberly E. Arnold and Matthew D. Pistilli. 2012. Course signals at Purdue. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*. ACM, New York, NY, USA, 267–270. DOI: https://doi.org/10.1145/2330601.2330666.

[3] Ryan S. Baker and Aaron Hawn. 2022. Algorithmic Bias in Education. *Int J Artif Intell Educ* 32, 4, 1052–1092. DOI: https://doi.org/10.1007/s40593-021-00285-9.

[4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, Cambridge.

[5] Patricia H. Blundon. 2016. Nonstandard dialect and educational achievement: Potential implications for first nations students. *Canadian Journal of Speech-Language Pathology and Audiology* 40, 3, 218–231.

[6] Ana Bravo-Moreno. 2010. Access and Transitions in Education. In *Changing Educational Landscapes*, Dimitris Mattheou, Ed. Springer Netherlands, Dordrecht, 107–118. DOI: https://doi.org/10.1007/978-90-481-8534-4_7.

[7] Julia Bredtmann, Sebastian Otten, and Christina Vonnahme. 2021. Linguistic diversity in the classroom, student achievement, and social integration. *Education Economics* 29, 2, 121–142. DOI: https://doi.org/10.1080/09645292.2020.1866499.

[8] Carol Campbell. 2021. Educational equity in Canada: the case of Ontario's strategies and actions to advance excellence and equity for students. *School Leadership & Management* 41, 4-5, 409–428. DOI: https://doi.org/10.1080/13632434.2019.1709165.

[9] Ismail Celik. 2023. Exploring the Determinants of Artificial Intelligence (AI) Literacy: Digital Divide, Computational Thinking, Cognitive Absorption. *Telematics and Informatics* 83, 102026. DOI: https://doi.org/10.1016/j.tele.2023.102026.

[10] Helen Chadwick, Louise Laverty, Robert Finnigan, Robert Elias, Ken Farrington, Fergus J. Caskey, and Sabine N. van der Veer. 2024. Engagement With Digital Health Technologies Among Older People Living in Socially Deprived Areas: Qualitative Study of Influencing Factors. *JMIR formative research* 8, e60483. DOI: https://doi.org/10.2196/60483.

[11] Hohsung Choe and Youngjoo Seo. 2021. Negotiating teacher identity: Experiences of Black teachers of English in Korean ELT. *English Today* 37, 3, 148–155. DOI: https://doi.org/10.1017/S0266078419000531.

[12] Patricia Hill Collins. 2015. Intersectionality's Definitional Dilemmas. *Annu. Rev. Sociol.* 41, 1, 1–20. DOI: https://doi.org/10.1146/annurev-soc-073014-112142.

[13] Patricia Hill Collins and Sirma Bilge. 2016. *Intersectionality*. Polity Press, Cambridge, UK.

[14] Kimberlé W. Crenshaw. 1989. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum* 1989, 139.

[15] Jenny L. Davis, Apryl Williams, and Michael W. Yang. 2021. Algorithmic reparation. *Big Data & Society* 8, 2, 205395172110448. DOI: https://doi.org/10.1177/20539517211044808.

[16] Marie Decker and Carmen Leicht-Scholten. 2024. „Intersektionale Algorithmen?“ Herausforderungen und Chancen der Operationalisierung von Intersektionalität. In *Journal Netzwerk Frauen- und Geschlechterforschung NRW*. DuEPublico: Duisburg-Essen Publications online, University of Duisburg-Essen, Germany, 92–99. DOI: https://doi.org/10.17185/duepublico/82760.

[17] Marie Decker, Laila Wegner, and Carmen Leicht-Scholten. 2024. Challenging "Subgroup Fairness": Towards Intersectional Algorithmic Fairness Based on Personas. In *Proceedings of the 3rd European Workshop on Algorithmic Fairness (EWAF24)*, Mattia Cerrato, Alesia Vallenas Coronel, Petra Ahrweiler, Michele Loi, Mykola Pechenizkiy and Aurelia Tamò-Larrieux, Eds. CEUR Workshop Proceedings, Aachen.

[18] Marie Decker, Laila Wegner, and Carmen Leicht-Scholten. 2025. Procedural fairness in algorithmic decision-making: the role of public engagement. *Ethics Inf Technol* 27, 1. DOI: https://doi.org/10.1007/s10676-024-09811-4.

[19] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

[20] James S. Etim. 2023. Improving Access and Educational Equity. In *Handbook of Research on Solutions for Equity and Social Justice in Education*, Siran Mukerji, Purnendu Tripathi, James Etim and Alice Etim, Eds. Advances in Educational Marketing, Administration, and Leadership. IGI Global, 1–18. DOI: https://doi.org/10.4018/978-1-7998-9678-4.ch001.

[21] Raymond E. Floyd and Richard H. Spencer. 2012. Education Challenge: ESL and Cultural Differences. *IEEE Potentials* 31, 4, 11–13. DOI: https://doi.org/10.1109/MPOT.2012.2195218.

[22] Usman Gohar and Lu Cheng. 2023. A Survey on Intersectional Fairness in Machine Learning: Notions, Mitigation, and Challenges. *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23). DOI: https://doi.org/*10.24963/ijcai.2023/742.

[23] Jette G. Hansen Edwards. 2017. Defining ‘native speaker’ in multilingual settings: English as a native language in Asia. *Journal of Multilingual and Multicultural Development* 38, 9, 757–771. DOI: https://doi.org/10.1080/01434632.2016.1257627.

[24] Lisa A. Harrison, Chiesha M. Stevens, Adrienne N. Monty, and Christine A. Coakley. 2006. The consequences of stereotype threat on the academic performance of White and non-White lower income college students. *Soc Psychol Educ* 9, 3, 341–357. DOI: https://doi.org/10.1007/s11218-005-5456-6.

[25] Scott Harrison, Renato Villano, Grace Lynch, and George Chen. 2016. Measuring financial implications of an early alert system. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16*. ACM Press, New York, New York, USA, 241–248. DOI: https://doi.org/10.1145/2883851.2883923.

[26] Anna L. Hoffmann. 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society* 22, 7, 900–915. DOI: https://doi.org/10.1080/1369118X.2019.1573912.

[27] Sandeep M. Jayaprakash, Erik W. Moody, Eitel J. Lauría, James R. Regan, and Joshua D. Baron. 2014. Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. *Learning Analytics* 1, 1, 6–47. DOI: https://doi.org/10.18608/jla.2014.11.3.

[28] Weijie Jiang and Zachary A. Pardos. 2021. Towards Equity and Algorithmic Fairness in Student Grade Prediction. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York, NY, USA, 608–617. DOI: https://doi.org/10.1145/3461702.3462623.

[29] Yang Jiang, Jiangang Hao, Michael Fauss, and Chen Li. 2024. Detecting ChatGPT-generated essays in a large-scale writing assessment: Is there a bias against non-native English speakers? *Computers & Education* 217, 105070. DOI: https://doi.org/10.1016/j.compedu.2024.105070.

[30] Niki Kilbertus, Mateo R. Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, 656–666.

[31] Youjin Kong. 2022. Are "Intersectionally Fair" AI Algorithms Really Fair to Women of Color? A Philosophical Analysis. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 485–494. DOI: https://doi.org/10.1145/3531146.3533114.

[32] Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. GPT detectors are biased against non-native English writers. *Patterns (New York, N.Y.)* 4, 7, 100779. DOI: https://doi.org/10.1016/j.patter.2023.100779.

[33] Michelle Maroto, David Pettinicchio, and Andrew C. Patterson. 2019. Hierarchies of Categorical Disadvantage: Economic Insecurity at the Intersection of Disability, Gender, and Race. *Gender & Society* 33, 1, 64–93. DOI: https://doi.org/10.1177/0891243218794648.

[34] Zari McFadden and Lauren Alvarez. 2024. Performative Ethics From Within the Ivory Tower: How CS Practitioners Uphold Systems of Oppression. *Journal of Artificial Intelligence Research*.

[35] Ursula Moffitt, Linda P. Juang, and Moin Syed. 2020. Intersectionality and Youth Identity Development Research in Europe. *Frontiers in psychology* 11, 78. DOI: https://doi.org/10.3389/fpsyg.2020.00078.

[36] José Moreno Montoya, Alvaro J. Idrovo, and Silvia M. Ballesteros. 2023. Short-term effects of COVID-19 pandemic on academic achievement among Colombian adolescents. *Rev. salud pública* 25, 2, 1–7. DOI: https://doi.org/10.15446/rsap.v25n2.106463.

[37] Sasha Nikolic, Carolyn Sandison, Rezwanul Haque, Scott Daniel, Sarah Grundy, Marina Belkina, Sarah Lyden, Ghulam M. Hassan, and Peter Neal. 2024. ChatGPT, Copilot, Gemini, SciSpace and Wolfram versus higher education assessments: an updated multi-institutional study of the academic integrity impacts of Generative Artificial Intelligence (GenAI) on assessment, teaching and learning in engineering. *Australasian Journal of Engineering Education* 29, 2, 126–153. DOI: https://doi.org/10.1080/22054952.2024.2372154.

[38] Anaelia Ovalle, Arjun Subramonian, Vagrant Gautam, Gilbert Gee, and Kai-Wei Chang. 2023. Factoring the Matrix of Domination: A Critical Review and Reimagination of Intersectionality in AI Fairness. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York, NY, USA, 496–511. DOI: https://doi.org/10.1145/3600211.3604705.

[39] Jessica D. Remedios and A.Chyei Vinluan. 2024. Intersectionality in social cognition: Stereotyping across interlocking social categories. *The Oxford Handbook of Social Cognition, Second Edition*.

[40] André Renz and Romy Hilbig. 2020. Prerequisites for artificial intelligence in further education: identification of drivers, barriers, and business models of educational technology companies. *Int J Educ Technol High Educ* 17, 1. DOI: https://doi.org/10.1186/s41239-020-00193-3.

[41] Catherine Riegle-Crumb, Sarah B. Kyte, and Karisma Morton. 2018. Gender and Racial/Ethnic Differences in Educational Outcomes: Examining Patterns, Explanations, and New Directions for Research. In *Handbook of the Sociology of Education in the 21st Century*, Barbara Schneider, Ed. Handbooks of Sociology and Social Research. Springer International Publishing, Cham, 131–152. DOI: https://doi.org/10.1007/978-3-319-76694-2_6.

[42] Richard Rothstein. 2018. To improve student learning: Ask the right questions. *The Next Agenda: Blueprint for a New Progressive Movement. DOI: https://doi.org/*10.4324/9780429496110.

[43] Sokip. 2020. Overcoming the problem of learning foreign language skills in the classroom. *Universal Journal of Educational Research* 8, 2, 723–729. DOI: https://doi.org/10.13189/ujer.2020.080246.

[44] Zainab Suntai and Susanny J. Beltran. 2023. The Intersectional Impact of Race/Ethnicity and Sex on Access to Technology Among Older Adults. *The Gerontologist* 63, 7, 1162–1171. DOI: https://doi.org/10.1093/geront/gnac178.

[45] Ceciley Zhang, Ronald E. Rice, and Laurent H. Wang. 2024. College students' literacy, ChatGPT activities, educational outcomes, and trust from a digital divide perspective. *New Media & Society. DOI: https://doi.org/*10.1177/14614448241301741.