Fairness in machine learning: A design research study for secondary education

Clara Müller, Katharina Bata, Martin Frank and Jasmin Hörter
Karlsruhe Institute of Technology, Germany
clara.mueller@kit.edu

This conceptual paper presents a design research study that explores fairness in machine learning (ML) as an interdisciplinary learning opportunity for secondary education. Drawing on the societal relevance of algorithmic decision-making systems, the study emphasizes the importance of integrating technical and ethical perspectives within a cohesive teaching-learning arrangement – an approach that is still rarely implemented in practice. The paper provides an overview of fairness definitions and algorithmic intervention strategies, alongside a review of relevant educational research on ML and fairness in both school and higher education contexts. It outlines the methodological foundations of the design research study and introduces preliminary ideas for a prototypical teaching-learning arrangement, accompanied by guiding research questions that frame the study.

INTRODUCTION

Machine learning (ML) systems are increasingly embedded in everyday decision-making processes with far-reaching societal implications. While these systems offer considerable benefits, they also risk reproducing or amplifying social biases (Mehrabi et al., 2022). Notable examples include automated hiring systems and loan approval algorithms that systematically disadvantage individuals from certain demographic groups (Ajunwa & Greene, 2019; Hardt et al., 2016).

In response to such challenges, fairness in ML has emerged as a central concern within an interdisciplinary research landscape spanning mathematics, statistics, computer science, ethics, and law. One of the core challenges in this field lies in making fairness measurable and actionable within the ML pipeline – a task that depends on the development and use of mathematical models. This includes the development of fairness definitions and algorithmic intervention strategies, alongside critical discussions about trade-offs – both between competing fairness notions and between fairness and model performance (Caton & Haas, 2024). ML constitutes the foundation of many AI applications and plays a pivotal role in algorithmic decision-making, making fairness in ML an essential dimension of broader debates on AI ethics.

Given the increasing societal impact of AI technologies, there is a growing recognition of the importance of empowering students to critically engage with and evaluate these systems. This is reflected in the development of various AI literacy frameworks and educational resources targeting K-12 education (Long & Magerko, 2020; Touretzky et al., 2019, 2023; UNESCO, 2024). Yet, most existing educational approaches focus either on the technical foundations or, to a lesser extent, on ethical and societal implications. However, especially in the context of fairness in ML, the integration of both perspectives is crucial.

This research project addresses this integration by presenting a design research framework that combines ethical reflection with computational and mathematical modeling. It aims to develop and empirically investigate a teaching-learning arrangement on fairness in ML tailored to secondary education.

FAIRNESS IN MACHINE LEARNING AS LEARNING CONTENT

Technical Background

This work focuses on supervised ML systems, which rely on training models using labeled data. The design process involves selecting a model class – such as decision trees or support vector machines – and defining a loss function that quantifies the discrepancy between the model's predictions and the actual outcomes. The training process then aims to minimize the loss function, thereby optimizing the model's predictive performance on the training data. However, even a well-fitted model can exhibit poor prediction accuracy on unseen data. Prediction errors – such as false positives or false negatives – may result from limitations in the dataset, the choice of the model class, or the choice of the loss function. Fairness becomes a concern when such errors, or more broadly the distribution of model outcomes, lead to systematic disadvantages for certain groups or individuals.

In: Birk, L., Loth, G., Jotzo, L., Binder, K., & Frischemeier, D. (Eds.) (2026). *Statistics and Data Science Education in STEAM. Proceedings of the Satellite Conference of the International Association for Statistical Education (IASE), September/October 2025, Münster, Germany*.

Recent research on fairness in ML has focused on the formalization of fairness through quantitative definitions and the development of intervention strategies to mitigate algorithmic bias. Numerous statistical definitions have been proposed to quantify fairness in algorithmic decision-making. A common approach involves comparing model outcomes across different subgroups defined by protected or sensitive attributes such as age, gender, or ethnicity (Caton & Haas, 2024). Among the most frequently discussed definitions are statistical parity, which requires equal rates of positive predictions across groups (Dwork et al., 2012), equal opportunity, which ensures equal true positive rates across groups (Hardt et al., 2016), and equalized odds, which requires not only equal true positive rates but also equal false positive rates across groups (Hardt et al., 2016). Another important fairness criterion is test fairness, which requires that individuals with the same predicted score have the same likelihood of belonging to the positive class, regardless of group membership (Chouldechova, 2017). There is no consensus on which of the numerous definitions is the best; many cannot be satisfied simultaneously (Kleinberg et al., 2017). Each definition reflects a different normative perspective on what it means for a model to be fair and must be selected with regard to the specific context (Verma & Rubin, 2018).

To operationalize fairness, various intervention strategies have been developed, targeting different stages of the ML pipeline (Caton & Haas, 2024): Pre-processing methods aim to reduce or remove embedded biases by modifying the input data prior to training, for instance by assigning weights to data points. In-processing methods integrate fairness objectives directly into the model training process by modifying the learning algorithm itself. Post-processing methods are applied after training and include approaches that transform model outputs to satisfy fairness constraints, for example by adjusting decision thresholds. Each strategy entails specific assumptions and limitations and requires balancing competing objectives – most notably, the trade-off between fairness and predictive accuracy (Liu & Vicente, 2022). Consequently, selecting appropriate fairness definitions and intervention strategies involves not only technical considerations but also value-laden decisions about how fairness is defined, interpreted, and prioritized within the specific context of application.

Learning Potential

Framing fairness in ML as a topic for secondary education, particularly within statistics and data science education, offers rich opportunities for interdisciplinary and socially meaningful learning. The concept provides a tangible and relevant context in which students can engage with key statistical ideas – including sampling, conditional probability, distributions, and statistical inference – while also fostering critical reflection on the ethical implications of data-driven technologies. A key analytical tool in this context is the confusion matrix, a specific type of contingency matrix that organizes model predictions into the following categories: true positives, false positives, true negatives, and false negatives. Working with these categories allows students to recognize that ML models are inherently subject to error. More importantly, it invites reflection on how different types of errors can carry unequal consequences across contexts – such as denying a qualified applicant a loan versus approving an unqualified applicant – thus linking statistical reasoning with broader questions of fairness and social impact.

In recent years, a growing number of K-12 initiatives have introduced students to the technical foundations of ML, focusing on what ML systems are and how they are created. Research indicates that learners across different age groups are capable of understanding foundational ML concepts and practices (Marques et al., 2020; Martins & Gresse von Wangenheim, 2023). Alongside these technical efforts, an emerging field of research has begun to explore teaching about the ethical and societal dimensions of ML systems (Marques et al., 2020). These initiatives aim to foster students' awareness of issues such as bias, discrimination, transparency, and accountability (Akgun & Greenhow, 2022; Ali et al., 2019). Touretzky et al. (2019) emphasize the importance of teaching students to recognize the potential of AI to both benefit and harm society as a central educational objective. More recently, Dabbagh et al. (2025) have called for AI ethics to become a mandatory part of school education, as being key to equipping students to responsibly navigate an AI-driven future.

Despite these advances, few educational approaches integrate both technical and ethical perspectives in a coherent way (Bilstrup et al., 2020; Kaspersen et al., 2021; Skinner et al., 2020; Williams et al., 2023). In particular, the topic of fairness in ML remains underrepresented, especially in

ways that explicitly engage with its technical foundations, such as fairness definitions and intervention strategies. Yet, this integration of both perspectives is essential: meaningful learning about fairness in ML requires students to understand not only how fairness can be formalized, measured, and operationalized but also to critically reflect on the closely intertwined ethical implications of the technical decisions. As an early example of such integration at the secondary level, Schönbrodt et al. (in press) describe a classroom activity on fairness in data-driven, algorithmic decision-making in the context of credit granting, in which students analyze decision boundaries for two groups and engage with stacked dot plots and related statistical measures.

At university level, various approaches in ML and data science education combine technical instruction with ethical reflection (Baumer et al., 2020; Borenstein & Howard, 2021; Grosz et al., 2019; Saltz et al., 2019; Skirpan et al., 2018). These approaches offer important insights for adaptation to secondary education. For example, studies on teaching university students about fairness in ML have shown that visualizations and interactive tools can improve students' understanding of fairness concepts (Mashhadi et al., 2022; Yan et al., 2024).

The educational potential of fairness extends beyond ML-specific education. The concept of fairness itself has been identified as a core educational topic, fundamental to fostering an inclusive and equitable society (Li et al., 2016; Skovsmose, 2023). Prior research in K-12 mathematics education has explored fairness through mathematical modeling contexts such as distributing shared resources or ranking performances. For example, Ärlebäck and Frejd (2025) demonstrate how secondary students tasked with ranking heptathlon results developed and reflected on different mathematical models of fairness, grounded in varying assumptions and priorities.

In the context of ML, fairness definitions can be understood as instances of prescriptive modeling – modeling that does not primarily aim to describe a phenomenon of the world, but rather to design, prescribe, organize or structure certain aspects of it (Niss, 2015). This perspective creates opportunities for engaging students in meta-validation (Niss, 2015). It encourages students to critically examine the consequences of modeling outcomes for affected individuals and groups, to compare the chosen model with alternative approaches, and to consider how the model might be adapted under changing initial conditions.

DESIGN RESEARCH ON FAIRNESS IN MACHINE LEARNING

Ideas and goals

Building on its interdisciplinary and educational potential, fairness in ML emerges as a rich and topical subject for secondary education, offering both mathematical depth and social relevance. Due to its complexity and novelty, it requires targeted research into how it can be effectively taught and learned. This project addresses this need through a design research approach guided by the overarching question: *How can the subject of fairness in ML be effectively taught and learned in secondary education, considering both ethical and technical aspects?* To explore this question, the project is structured along two closely interwoven strands: design and empirical research.

The first strand focuses on the development of a concrete teaching-learning arrangement on fairness in ML, specifically designed for secondary education. The primary objective is to develop a classroom-tested module that can serve as a starting point for broader classroom implementation. Beyond this practical aim, the design process also contributes to the theoretical discourse in subject-specific educational research. By specifying and structuring the learning content related to fairness in ML and by developing or adapting design principles, the project aims to generate theoretical insights that can inform future instructional designs.

Central research questions in specifying and structuring the learning content on fairness in ML include:

- What are the relevant learning objectives?
- What are the relevant mathematical content elements, and how are they inherently connected?
- How can these elements be structured into a coherent intended learning trajectory?

The various fairness definitions and intervention strategies outlined above – particularly in relation to mathematical modeling – offer rich opportunities for the development of learning materials. Their connection to ethical considerations and the potential for reflecting on decision-making processes

further expands this potential. A corresponding learning trajectory could, for instance, encompass the following key elements:

- Developing a conceptual understanding of fairness: Learners begin by reflecting on their everyday conceptions of fairness, shaped by personal experiences and sociocultural contexts. These informal understandings serve as an entry point for the introduction of formal, mathematically grounded fairness definitions commonly used in ML, such as statistical parity, equal opportunity, and equalized odds. Students are encouraged to critically compare their intuitive concepts with formal models by reflecting on the normative assumptions embedded in each definition.
- Analyzing fairness in real-world data: Building on this conceptual foundation, students apply selected fairness criteria to real-world datasets. They analyze the distribution of predictions across sensitive groups, assess fairness-related discrepancies, and interpret the societal implications of their findings. This hands-on exploration is designed to bridge the gap between abstract fairness concepts and their real-world implications.
- Exploring fairness intervention strategies: Students are introduced to selected fairness intervention techniques. Through practical experimentation, learners explore how interventions affect fairness outcomes and model behavior.
- Discussing trade-offs and competing priorities: Structured reflection phases are embedded throughout the learning sequence to support the critical engagement with societal implications of technical design decisions. Students discuss inherent trade-offs between competing fairness definitions and the potential impact of fairness interventions on model performance and transparency. These dialogues aim to foster integrated reasoning that bridges technical analysis with ethical considerations involved in operationalizing fairness in ML systems.

The design work forms the foundation for the second strand of the project, the empirical research. This phase aims to evaluate the effectiveness of the teaching-learning arrangement in initiating and supporting the intended learning processes and to derive evidence-based insights into how students engage with both the technical and ethical aspects of fairness in ML. It seeks to develop empirically grounded local theories of the teaching and learning processes associated with the developed teaching-learning arrangement.

Central research questions concerning the specific teaching-learning arrangement include:

- To what extent does the developed teaching-learning arrangement effectively initiate the intended learning processes, and how well do these processes align with the intended learning trajectories?
- What are content-specific learning processes, including typical learning pathways, difficulties, and obstacles?
- What are content-specific teaching processes, including typical conditions and effects?

Methods

To address the two strands of design and empirical research, the project adopts a design research methodology that integrates the development of teaching-learning arrangements with subject-specific educational research. The study is guided by the Dortmund model of topic-specific design research proposed by Prediger et al. (2012), which combines a content-focused and process-focused approach. This dual orientation connects, on the one hand, to Wittmann's (1995) conceptualization of mathematics education as a 'design science' that structures subject matter from the learner's perspective. On the other hand, it draws on Gravemeijer and Cobb's (2006) notion of design research as a method for investigating learning processes.

Based on Prediger et al. (2012), the design research is structured into iterative cycles across four intertwined working areas:

1. *Specification and structuring of the learning content* – Core mathematical concepts (e.g., confusion matrix, conditional probability, statistical parity) are identified and paired with authentic, real-world contexts. Example datasets include the *German Credit* dataset (Dua &

Graff, 2017) to explore gender disparities in credit decisions, and the *COMPAS* dataset (Larson et al., 2016) to explore fairness issues in recidivism prediction.

2. *Development of the teaching-learning arrangement* – The teaching-learning arrangement is developed within the framework of CAMMP (Computational and Mathematical Modeling Program; Frank & Roeckerath, 2022). CAMMP aims to teach secondary students the relevance of mathematics for understanding socially relevant topics – such as fairness in ML. The developed arrangement is intended for CAMMP Days – half- or full-day workshops conducted with entire school classes. It follows CAMMP's design principles, namely working on real, relevant problems through mathematical modeling and the use of digital tools (Hofmann et al., in press). The learning activities center on interactive, digital worksheets in Jupyter Notebooks, which combine explanatory texts, visuals, and code. Automated feedback and differentiated scaffolding ensure accessibility for learners without prior programming experience.

   Preliminary example of learning material: Reflection on the trade-off between fairness and model performance
   Visual representations such as Figure 1 can serve as an entry point for investigating how changes in fairness may affect model performance. Students may be tasked with interpreting the figure, identifying patterns, and discussing potential explanations. The activity is designed to prompt reflection on the role of context: for example, in predicting criminal recidivism, a trade-off may arise between accuracy – which in this setting is linked to public safety – and fairness toward all individuals involved (Corbett-Davies et al., 2017). In such contexts, prioritizing high accuracy could be considered appropriate. Conversely, in allocation scenarios such as the distribution of scholarships or training opportunities, a higher degree of fairness may be preferable, even if this entails some reduction in predictive performance.

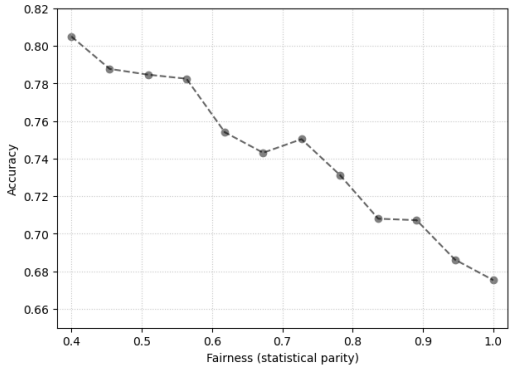

Figure 1. Relationship between fairness (statistical parity-based metric with 1 indicating maximal fairness) and predictive performance (accuracy) for a toy dataset.

3. *Conducting and analyzing design experiments* – The developed teaching-learning arrangement is tested through iterative design experiments in laboratory and classroom settings. Data sources include student work artifacts (e.g., completed notebooks), classroom observations, and semi-structured interviews with students and teachers, enabling analysis of how learning processes unfold in practice.
4. *Development of local theories* – Findings from these experiments are synthesized into local theories about teaching and learning processes related to fairness in ML. These theories describe typical learning pathways, challenges, and effective instructional strategies for fostering technical-ethical reasoning about fairness in ML.

DISCUSSION AND OUTLOOK

This paper has outlined different reasons for fairness in ML holding substantial potential as an interdisciplinary and socially relevant topic for secondary education. In line with existing approaches,

it is argued that the joint development of the mathematical foundations and the ethical perspective is not only mutually enriching (Williams et al., 2023), but also both feasible and pedagogically valuable within school contexts. Building on this premise, the paper presents a design-based research approach aimed at systematically developing and investigating a corresponding teaching-learning arrangement. In implementing the project, particular care must be taken in selecting from the wide range of possible content and instructional strategies to ensure the material is neither overwhelming for learners nor lacking in depth with respect to technical and ethical dimensions.

The next steps of the project follow the outlined methodological approach (Prediger et al., 2012) and include the specification and structuring of learning content, as well as the development of a prototypical teaching-learning arrangement. This entails the careful selection of a relevant real-world problem and an accompanying dataset. By the time of the conference in fall 2025, initial task designs and sample instructional sequences will have been developed. The presentation will therefore be enriched with preliminary examples of learning material.

REFERENCES

Ajunwa, I., & Greene, D. (2019). Platforms at work: Automated hiring platforms and other new intermediaries in the organization of work. In K. Vallas & A. Kovalainen (Eds.), *Work and labor in the digital age* (Vol. 33, pp. 61–91). Emerald Publishing. https://doi.org/10.1108/S0277-283320190000033005

Akgun, S., & Greenhow, C. (2022). Artificial intelligence in education: Addressing ethical challenges in K-12 settings. *AI and Ethics*, *2*(3), 431–440. https://doi.org/10.1007/s43681-021-00096-7

Ali, S., Payne, B. H., Williams, R., Park, H. W., & Breazeal, C. (2019). Constructionism, ethics, and creativity: Developing primary and middle school artificial intelligence education. In *Proceedings of the International Workshop on Education in Artificial Intelligence K-12 (EDUAI'19), IJCAI 2019* (pp. 1–4). IJCAI https://robots.media.mit.edu/wp-content/uploads/sites/7/2019/08/Constructionism__Ethics__and_Creativity.pdf

Ärlebäck, J. B., & Frejd, P. (2025). Exploring Swedish secondary students' mathematical models of fairness in ranking real-life, multi-faceted heptathlon outcomes. *ZDM – Mathematics Education, 57*, 229–244. https://doi.org/10.1007/s11858-025-01678-z

Baumer, B. S., Garcia, R. L., Kim, A. Y., Kinnaird, K. M., & Ott, M. Q. (2020). *Integrating data science ethics into an undergraduate major: A case study*. Smith College, Statistical and Data Sciences. https://scholarworks.smith.edu/sds_facpubs/20

Bilstrup, K.-E. K., Kaspersen, M. H., & Petersen, M. G. (2020). Staging reflections on ethical dilemmas in machine learning: A card-based design workshop for high school students. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, (pp. 1211–1222). https://doi.org/10.1145/3357236.3395558

Borenstein, J., & Howard, A. (2021). Emerging challenges in AI and the need for AI ethics education. *AI and Ethics*, *1*(1), 61–65. https://doi.org/10.1007/s43681-020-00002-7

Caton, S., & Haas, C. (2024). Fairness in machine learning: A survey. *ACM Computing Surveys*, *56*(7), Article 166. https://doi.org/10.1145/3616865

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 797–806). https://doi.org/10.1145/3097983.3098095

Dabbagh, H., Earp, B. D., Mann, S. P., Plozza, M., Salloch, S., & Savulescu, J. (2025). AI ethics should be mandatory for schoolchildren. *AI and Ethics*, *5*(1), 87–92. https://doi.org/10.1007/s43681-024-00462-1

Dua, D., & Graff, C. (2017). *UCI machine learning repository* [Data set]. University of California, Irvine. http://archive.ics.uci.edu/ml

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, (pp. 214–226). https://doi.org/10.1145/2090236.2090255

Frank, M., & Roeckerath, C. (Eds.). (2022). *Neue Materialien für einen realitätsbezogenen Mathematikunterricht 9: Mathematische Modellierung in interdisziplinären Projekten für die Sekundarstufe (ab Klasse 9)*. [New materials for reality-based mathematics education 9: Mathematical modeling in interdisciplinary projects for secondary education (from grade 9 onwards)] ISTRON-Schriftenreihe. Springer Spektrum. https://doi.org/10.1007/978-3-662-63647-3

Gravemeijer, K. P. E., & Cobb, P. (2006). Design research from a learning design perspective. In J. Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), *Educational Design Research* (pp. 45–85). Taylor & Francis.

Grosz, B. J., Grant, D. G., Vredenburgh, K., Behrends, J., Hu, L., Simmons, A., & Waldo, J. (2019). Embedded EthiCS: Integrating ethics across CS education. *Communications of the ACM*, *62*(8), 54–61. https://doi.org/10.1145/3330794

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, (Vol. 29). https://doi.org/10.48550/arXiv.1610.02413

Hofmann, S., Bata, K., Schönbrodt, S., Kindler, S., Frank, M., Müller, C., Eckert, J., Hörter, J., Büsing, C., & Stamm, B. (in press). Mathematical modeling in action: Design principles for mathematical modeling days. *PAMM*.

Kaspersen, M. H., Bilstrup, K.-E. K., Van Mechelen, M., Hjorth, A., Bouvin, N. O., & Petersen, M. G. (2021). VotestratesML: A high school learning tool for exploring machine learning and its societal implications. In *FabLearn Europe / MakeEd 2021 - An International Conference on Computing, Design and Making in Education*, 1–10. https://doi.org/10.1145/3466725.3466728

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the International Conference on Theory of Computing (ITCS)* 2017 (Vol. 67, Article 43, pp. 43:1–43:23). https://doi.org/10.4230/LIPICS.ITCS.2017.43

Larson, S., Mattu, L., Kirchner, L., & Angwin, J. (2016). *ProPublica COMPAS analysis data set* [Data set]. GitHub. https://github.com/propublica/compas-analysis

Li, J., Wang, W., Yu, J., & Zhu, L. (2016). Young children's development of fairness preference. *Frontiers in Psychology*, *7,* Article 1274. https://doi.org/10.3389/fpsyg.2016.01274

Liu, S., & Vicente, L. N. (2022). Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. *Computational Management Science*, *19*, 513–537. https://doi.org/10.1007/s10287-022-00425-z

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, (pp. 1–16). https://doi.org/10.1145/3313831.3376727

Marques, L. S., Gresse von Wangenheim, C., & Hauck, J. C. R. (2020). Teaching machine learning in school: A systematic mapping of the state of the art. *Informatics in Education*, *19*(2), 283–321. https://doi.org/10.15388/infedu.2020.14

Martins, R. M., & Gresse von Wangenheim, C. (2023). Findings on teaching machine learning in high school: A ten-year systematic literature review. *Informatics in Education*, *22*(3), 421–440. https://doi.org/10.15388/infedu.2023.18

Mashhadi, A., Zolyomi, A., & Quedado, J. (2022). A case study of integrating fairness visualization tools in machine learning education. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–7. https://doi.org/10.1145/3491101.3503568

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), Article 115. https://doi.org/10.1145/3457607

Niss, M. (2015). Prescriptive modelling—Challenges and opportunities. In G. A. Stillman, W. Blum, & M. S. Biembengut (Eds.), *Mathematical modelling in education research and practice: Cultural, social, and cognitive influences* (pp. 67–79). Springer.

Prediger, S., Link, M., Hinz, R., Hußmann, S., Thiele, J., & Ralle, B. (2012). Lehr-Lernprozesse initiieren und erforschen – Fachdidaktische Entwicklungsforschung im Dortmunder Modell. *Mathematisch-Naturwissenschaftlicher Unterricht*, 65(8), 452–457.

Saltz, J., Skirpan, M., Fiesler, C., Gorelick, M., Yeh, T., Heckman, R., Dewar, N., & Beard, N. (2019). Integrating ethics within machine learning courses. *ACM Transactions on Computing Education*, *19*(4), 32:1–32:26. https://doi.org/10.1145/3341164

Schönbrodt, S., Schneider, S., Podworny, S., & Camminady, T. (in press). A learning activity on fairness in data-driven algorithmic decision-making systems. *Teaching Statistics*.

Skinner, Z., Brown, S., & Walsh, G. (2020). Children of color's perceptions of fairness in AI: An exploration of equitable and inclusive co-design. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–8. https://doi.org/10.1145/3334480.3382901

Skirpan, M., Beard, N., Bhaduri, S., Fiesler, C., & Yeh, T. (2018). Ethics education in context: A case study of novel ethics activities for the CS classroom. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, 940–945. https://doi.org/10.1145/3159450.3159573

Skovsmose, O. (2023). *Critical Mathematics Education*. Springer International Publishing. https://doi.org/10.1007/978-3-031-26242-5

Touretzky, D., Gardner-McCune, C., & Seehorn, D. (2023). Machine learning and the five big ideas in AI. *International Journal of Artificial Intelligence in Education*, *33*(2), 233–266. https://doi.org/10.1007/s40593-022-00314-1

Touretzky, D., Gardner-McCune, C., Martin, F., & Seehorn, D. (2019). Envisioning AI for K-12: What should every child know about AI? *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(1), Article 9795. https://doi.org/10.1609/aaai.v33i01.33019795

UNESCO. (2024). *AI competency framework for students*. https://doi.org/10.54675/JKJB9835

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, 1–7. https://doi.org/10.1145/3194770.3194776

Williams, R., Ali, S., Devasia, N., DiPaola, D., Hong, J., Kaputsos, S. P., Jordan, B., & Breazeal, C. (2023). AI + Ethics curricula for middle school youth: Lessons learned from three project-based curricula. *International Journal of Artificial Intelligence in Education*, *33*(2), 325–383. https://doi.org/10.1007/s40593-022-00298-y

Wittmann, E. Ch. (1995). Mathematics education as a design science. *Educational Studies in Mathematics*, *29*(4), 355–374. https://doi.org/10.1007/BF01273911

Yan, X., Zhou, Y., Mishra, A., Mishra, H., & Wang, B. (2024). Exploring visualization for fairness in AI education. In *2024 IEEE 17th Pacific Visualization Conference (PacificVis)*, 1–10. https://doi.org/10.1109/PacificVis60374.2024.00010