A design research project on fairness in data-driven algorithmic decision-making

Sarah Schönbrodt[1] and Steffen Schneider[2]
[1]Paris Lodron University Salzburg, Austria; [2]KI macht Schule, Germany
sarah.schoenbrodt@plus.ac.at

Data-driven algorithmic decision-making systems, including many AI technologies, are ubiquitous in today's society. Early engagement with these systems and the promotion of critical statistical literacy are therefore essential. This involves not only fostering a basic understanding of how such systems work, but also addressing their broader social impact. As part of a design research project, we developed a learning activity that enables upper secondary students and pre-service teachers to explore issues of fairness in the context of automated credit granting. In this paper, we present the design of the activity, outline the intended learning trajectory, and report on initial implementations with pre-service teachers. We also discuss preliminary findings from a qualitative analysis of students' fairness-related arguments.

INTRODUCTION

Data-driven algorithmic decision-making systems (ADMS) play a crucial role in our society. They are applied in various areas such as credit granting, personalized online advertising, or the selection of job applicants. However, significant problems can arise in the implementation of such systems, particularly in the form of discrimination against certain groups (O'Neil, 2016; Hardt et al., 2016; Orwat, 2019). This discrimination can be subtle and unintended but have far-reaching consequences. Although fairness and non-discrimination in ADMS have been discussed for decades (Caton & Haas, 2024), the topic has become particularly relevant with the rise of AI and machine learning.

The increasing importance of data-driven ADMS necessitates addressing these topics in high schools. Students should learn to critically engage with data-driven ADMS and reflect upon possible discriminatory effects (Hilger & Büscher, in press). This leads to the question of how learning activities can be designed that allow students to develop and discuss fairness approaches in the development of ADMS. Additionally, it is interesting to examine what statistical argumentative bases students use for their fairness approaches.

THEORETICAL CONSIDERATIONS

We first outline definitions of algorithmic fairness and review prominent approaches from the scientific literature. Next, we consider the concept of critical statistical literacy to evaluate how the learning activity addresses its core components.

Approaches and statistical measures of algorithmic fairness

One approach to construe fairness in the context of ADMS discussed by Barocas et al. (2023) is fairness as the absence of discrimination. This approach studies how personal characteristics of an individual influence decision outcomes. Here, discrimination means that someone is treated wrongfully because of their group membership (e.g., gender, race). Discrimination is not a universal concept but is context-dependent, as it relates to ways in which people's lives are affected. It is about socially significant classifications that have served as the basis for unjustified and systematically disadvantageous treatment (Barocas et al., 2023, p. vi).

Numerous fairness approaches exist, each with distinct mathematical implementations. Here we briefly discuss two approaches using the problem setting mirrored in our learning activity: assigning credit applicants of two groups (e.g., male and female) to the classes "creditworthy" (positive prediction) and "not creditworthy" (negative prediction) based on their credit score. Since these scores are often derived from algorithms trained on historical data, they can encode statistical bias and – when used to decide on loan approval – lead to discrimination against certain groups.

How would a fair credit granting system treat individuals of both groups? Two (among many) approaches that are discussed in the literature (Hardt et al., 2016) are:

(I) Demographic parity: Ensure equal loan approval rate across groups. To implement this approach, the difference in *positive rates* in both groups is minimized.

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*.

(II) Equal opportunity: Guarantee equal approval rates for creditworthy individuals across groups. To implement this approach, the difference in *true positive rates (TPR)* in both groups is minimized.

Numerous other notions of fairness exist, requiring different statistical measures for their implementation (Pessach & Shmueli, 2020). It is crucial to recognize that no universally optimal fairness approach exists for all contexts. Research has even demonstrated that it is typically impossible to simultaneously satisfy multiple meaningful fairness approaches (Pessach & Shmueli, 2020). Consequently, selecting the appropriate fairness strategy becomes a value-based decision that must account for social, ethical, and legal considerations.

Fairness approaches can be integrated at various stages during the development of an ADMS. One possibility is to address fairness *before* the algorithm is developed, for instance by detecting and reducing systematic bias in the dataset (*pre-process phase*). Alternatively, fairness measures can be incorporated *during* the development of the algorithm itself, a stage often referred to as the training phase in machine learning. At this point, concepts of fairness and their mathematical formulations can be embedded directly into the training process, such as by adapting the objective function when training a neural network (*in-process phase)*. Finally, it is possible to address fairness *after* the algorithm has been developed by modifying its predictions. This *post-process* approach may involve adjusting parameters like decision thresholds to ensure that the ADMS produces fair outcomes for different groups (Pessach & Shmueli, 2020).

The learning activity described here centers on a post-processing approach to fairness. This focus offers pedagogical advantages. It circumvents the need to analyze complex, multidimensional data, as required in most pre-processing methods, and does not demand a deep understanding of the underlying algorithms, as is necessary for in-processing interventions. As a result, students can directly engage with core concepts of algorithmic fairness and observe the impact of different fairness approaches on system outcomes – without requiring advanced technical expertise.

Critical statistical literacy

In his framework for critical statistical literacy (CSL), Weiland (2017) proposes a closer integration of traditional notions of statistical literacy with concepts from the field of critical literacy. The term statistical literacy itself is not uniformly defined, as various frameworks describe its core elements and outline relevant skills. Weiland's CSL framework combines two central perspectives on statistical literacy. The *consumer perspective,* as emphasized by Gal (2006), involves making sense of statistical information (reading and interpreting statistics), while the *producer perspective* centers on enquiry, data production and the active creation and communication of statistical arguments as emphasized for example in Gould (2010) or in the PPDAC-Cycle by Wild and Pfannkuch (1999). The consumer perspective can be associated with "reading the world through statistics" and the producer perspective with "writing the world through statistics" (Weiland, 2017, p. 39). Weiland argues that the consumer and the producer perspective entail deeply intertwined practices that are essential for critical engagement. Thus, the CSL framework encompasses both the consumer and the producer role and links these two roles to questioning social structures and facilitating transformative action. This broader critical stance encourages students to use statistics not only to evaluate information, but also to question social injustices and to take action (Weiland, 2017).

The CSL framework includes examples of skills that students should acquire with respect to both the reading- and writing-component. The skills around reading include "making sense of language and statistical symbols systems and critiquing statistical information and data-based arguments encountered in diverse contexts to gain an awareness of the systemic structures at play in society" and "identifying and interrogating social structures which shape and are reinforced by data based arguments" (Weiland, 2017, p. 41). The writing-skills include "using statistical investigations to communicate statistical information and arguments in an effort to destabilize and reshape structures of injustice for a more just society" and "using statistical investigations to alleviate and resolve sociopolitical issues of injustice" (Weiland, 2017, p. 41).

The learning activity described here is designed to address both the reading and writing components of CSL. By doing so, it aims to foster the skills necessary for students to engage critically with data and statistics and to use statistical reasoning as a tool to take action for social change.

Design Research

Our work follows the design research approach by Gravemeijer and Cobb (2006). This approach seeks to meaningfully integrate the processes of *design* and *research* by focusing on the creation and iterative refinement of educational environments through cycles of design, testing, and reflective analysis in authentic classroom contexts. The aim is to develop both practical instructional solutions and empirically grounded theories of learning, thereby enhancing educational understanding and effectiveness.

In this project, we developed a learning activity centered on an interactive web app addressing the topic of algorithmic fairness in automated credit granting. While the design of the learning activity has been presented in detail in Schönbrodt et al. (2025), the present paper contributes the first systematic empirical evaluation, thereby providing insights into its implementation and impact.

The design of the web app is based on an interactive blog post by Wattenberg, Viégas, and Hardt (n.d.), which highlights key findings on algorithmic fairness by Hardt et al. (2016). We adapted the blog post to make the topic accessible to high school students. The web app is available in both German and English at https://kiwi.schule/fairness under a CC license.

In the following sections, we describe the design of the learning activity, briefly revisit its central steps, and present initial empirical insights from a design experiment conducted with pre-service mathematics teachers.

A LEARNING ACTIVITY ON FAIRNESS IN CREDIT GRANTING

The learning activity enables students to explore, analyze, and articulate different notions of fairness in developing a data-driven ADMS for credit granting. It builds on students' prior knowledge of frequency diagrams, specifically stacked dot plots, contingency tables, and relative proportions (Schönbrodt et al., 2025). The learning activity was designed for implementation with upper secondary students (i.e., ages 14–19 in Austria) and pre-service teachers.

Design principles

For the design of the learning activity, several design principles (DP) were implemented:

- *DP 1: Implementing the learning activity as a low-threshold, interactive web app accessible through any browser.* This DP was implemented to ensure maximum accessibility for learners. No advanced technical skills or even programming skills are needed to work with the web app. The app is mainly operated using interactive sliders.
- *DP 2: Using a context that does not lead to negative emotional engagement.* A less emotionally charged context was chosen to prevent students from strongly identifying with particular groups, which could otherwise lead to emotionally heated discussions within the classroom. More precisely, we focused on the fictitious residency of two subpopulations as the sensitive attribute. This DP aims to strike a balance between addressing controversial issues and alleviating concerns that teachers may have about managing negative emotions or intense debates.
- *DP 3: Implementing the development of fairness approaches as open-ended tasks.* This DP should ensure that students are not limited to trying out predefined fairness approaches; rather, they are encouraged to develop their own strategies and support their choices using statistical arguments. Comparisons with established approaches used in practice can subsequently be discussed in the classroom. However, the focus remains on the students' individual solutions.
- *DP 4: Addressing both the reading and writing component of CSL*. Students should not only interpret the given data and detect discrimination caused by the data-driven ADMS, but also take action by developing their own fairness approaches.

Intended learning trajectory

Here we briefly outline the main steps of the intended learning trajectory implemented within the web app. For each step we highlight which component of CSL (i.e., the reading or the writing component) is emphasized.

Step 1: Understanding the context of credit granting (Reading). The activity begins with a brief introduction to the problem of discrimination in data-driven ADMS, illustrated by several real-world historical cases and their impacts on different social groups, including a case concerning credit granting. Next, students are introduced to the goal of developing a fair automated credit granting system. For simplicity, it is assumed that the bank only offers loans of 1,000€. If the bank grants a loan to a person who repays it, the bank earns 300€. Vice versa, if the bank issues a loan to someone who defaults, it incurs a loss of 700€, assuming the bank is insured. Students are asked to calculate the bank's overall profit for varying numbers of defaulting and non-defaulting individuals.

Step 2: Exploring the data and the classifier (Reading). The dataset consists of 400 individuals, each with known historical repayment status. Additionally, each individual has been assigned a credit score ranging from 0 to 100, computed by a complex model to which students do not have access. Thus, for each individual the "true" (historical) creditworthiness and the computed credit score (an imperfect proxy for actual creditworthiness) are available. The data are visualized as a stacked dot plot (see Figure 1A). Based on this visualization, the bank can set a decision threshold: individuals with a score equal to or above the threshold are predicted to likely repay the loan (positive class) and would receive a loan, while those below the threshold are more likely to default (negative class) and would be denied a loan. As shown in Figure 1A, there is overlap between the creditworthy (dark blue) and not creditworthy (light blue) data points, meaning the bank will inevitably make classification errors. Students are tasked to explore the data, intuitively select a decision threshold based on the provided plot, and justify their choice.

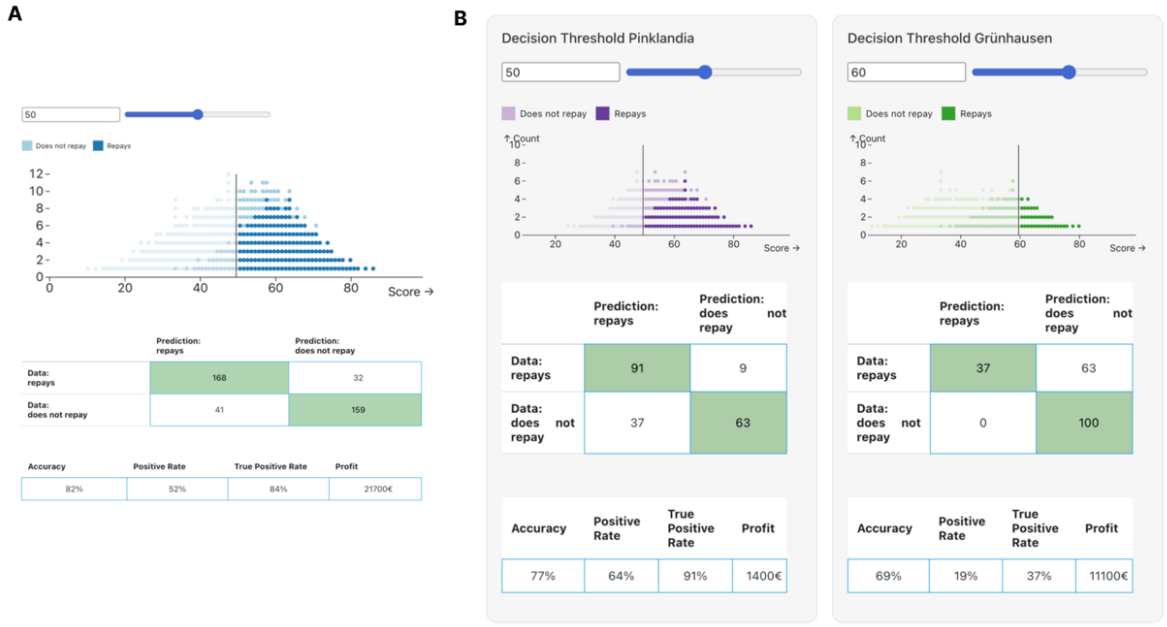

Figure 1. Dataset, decision threshold, contingency table, and statistical measures shown in the web app: A) for one population, and B) for two subpopulations.

Step 3: Computing and interpreting statistical measures (Reading and Writing). To properly evaluate the credit granting classifier, students are introduced to the contingency table as well as key statistical measures such as accuracy, positive rate, TPR, and total profit (see Figure 1A). Students calculate these statistical measures for a fixed decision threshold. Subsequently, students can dynamically adjust the threshold within the web app and select an optimal threshold for the population, examining the automatically computed statistical measures.

Step 4: Developing and implementing fairness approaches (Writing). Fairness becomes relevant when considering two different subgroups, distinguished by a sensitive attribute (e.g., ethnic origin, gender). In this activity, the (fictitious) attribute of residency is used. The dataset comprises 400 individuals, with 200 from the city of "Grünhausen" and 200 from the city of "Pinklandia." In each group, 100 individuals are creditworthy. Although the proportion of truly creditworthy individuals is

identical in both groups, the credit score distributions differ significantly (see Figure 1B), indicating that the credit score model produces some form of unequal treatment.

At this point, students address the central question: How should the decision thresholds for the two groups be chosen so that (a) the ADMS can be considered fair for both groups, and (b) the system remains profitable for the bank? Students can dynamically set thresholds for each group and, using the automatically calculated statistical measures, implement their own notions of fair decision-making (see Figure 1B). They are expected to justify their choices, explain why they consider the thresholds fair for both groups, and specify which statistical measures informed their decisions.

Step 5: Discussing the broader ethical implications (Writing). The fairness approaches derived in Step 4 are discussed in terms of their effects on different groups. Students critically reflect on the societal impacts of their chosen approach. They discuss how different fairness criteria influence specific subpopulations in dialogue with their peers. This discussion is not limited to technical results and should consider the perspectives of affected individuals, institutions, and society. The discussion centers on the complexities and trade-offs involved in defining and realizing fairness, including tensions between equity and profitability.

FIRST EMPIRICAL INSIGHTS

In the first design cycle, we conducted a design experiment with pre-service mathematics teachers at the University of Salzburg. The study aimed to address the following research question: *Which fairness approaches do students propose, and which statistical measures do they use to determine "fair" decision thresholds?*

Data Collection

A total of 22 pre-service teachers participated: 11 first-year students and 11 fifth-year students (in a six-year program). After an introductory session on ADMS generally and the development of an automatic credit granting system in particular, students independently completed the tasks embedded in the web app. The intervention lasted 90 minutes. We collected the students' written responses to all tasks. To answer the research question, we analyzed the written responses to two tasks from Step 4 of the learning trajectory:

- *Task 1*: Record the thresholds you selected for the two groups and explain why you consider these thresholds fair for both groups.
- *Task 2:* Which statistical measures did you use to determine the thresholds?

Initial Results

We analyzed the responses using an inductive-deductive qualitative content analysis. We categorized the responses into two main themes: (i) fairness approaches and (ii) statistical measures. Students expressed various notions of fairness, summarized in Table 1. They pursued different fairness approaches and, in some cases, explicitly referred to the statistical measures used.

Table 1. Fairness approaches, representative quotes, and number of students

Subcategories for "fairness approaches"	**Representative quote**	**Student counts**
As few false rejections as possible and high profit	"Because it is important that only those people who can pay back the loan receive it. The bank's profit probably also plays a role."	2
Equal standards for both groups and high profit	"It would be fair if the threshold was the same, but the bank also must ensure that it does not become insolvent, so I would recommend a threshold of 50 (for both). That way, not so many people are neglected, and the bank also makes a better profit."	3
Many people who receive a loan and high profit	"The bank's profit is not maximum but positive (32,400€) and many of the people get a loan."	1
High accuracy in both groups	"When looking at accuracy, the predictions most closely match the data, so bias can be avoided."	1

High accuracy in both groups and high profit	"I think 54 for Pink and 47 for Grünhausen are the fairest, as they have the highest accuracy (80%, 89%). Profit for bank also fits."	1
Treat groups independently of each other and maximize the profit	"Everyone is assessed individually for lending. Maximize profit for the bank."	3
High rate of correct decisions, and high rate of loans assigned to both groups	"For the groups of people, accuracy and positive rates are important and they want to feel that they are being treated fairly regarding these."	2
Profit maximization takes priority; fairness is irrelevant.	"Profit. Because the bank acts on this, that is the most important thing for it. The bank does not act humanly, but only profit- or money-oriented."	5
Not categorizable.	*Students did not explicitly describe their fairness approach or did not answer the task at all.*	4

To determine which statistical measures the students employed, we analyzed responses to *Task 2.* Among the 22 participants, 15 used total profit, while 13 referenced accuracy. Only five students used the TPR, and four the positive rate. Most often the students combined different measures, such as TPR and total profit.

To investigate the limited engagement with TPR and positive rate, we retrospectively analyzed students' performance in Step 3 of the learning activity. In this phase, students compute accuracy, positive rate, TPR, and total profit for a fixed decision threshold and interpret their relevance in contextual scenarios (e.g., identifying TPR as the appropriate measures for the question: *What percentage of creditworthy applicants receive loans?*). We observed that more than 90% of students correctly calculated accuracy (91%) and positive rate (95%), while only 59% succeeded with TPR. Similarly, only 77% correctly identified the TPR as relevant for the interpretation task.

Discussion

A surprisingly large number of students adopted the strategy, "Profit maximization takes priority; fairness is irrelevant." Several factors may account for this outcome. One notable limitation of our activity is that the example scenario may be too distant from the students' everyday realities; a more relatable context might prompt different decision-making. Future work could investigate whether alternative narratives influence students' strategies. Another possibility is that the learning activity does not sufficiently emphasize the importance of balancing perspectives of the different stakeholders, such as bank profitability *and* subpopulation equity. This raises questions about how educational designs can better position students to recognize, question, and address the societal impact of statistical models and decision thresholds, rather than focusing solely on quantitative optimization. One approach could be to explicitly place students in the role of independent policymakers to encourage strategies with a stronger societal focus.

Regarding the calculation and interpretation of the TPR, our findings indicate the need to refine explanatory materials. The lower rates of computational and interpretative success may have discouraged students from using this measure in subsequent fairness analyses. Incorporating interim discussion phases that focus on interpreting statistical measures might further strengthen students' conceptual understanding before they engage with more complex tasks, such as examining fairness approaches, profit-fairness trade-offs, and the broader societal and ethical implications of automated decision-making.

OUTLOOK

This design research project illustrates the interplay of statistical measures and the integration of ethical considerations in data-driven algorithmic decision-making. The learning activity allows students both to recognize discrimination in data-driven ADMS (consumer role) and to actively develop their own approaches to implementing fairness in such systems (producer role). Based on the findings from the first implementation, the learning activity will be revised, for example by reformulating the explanations on TPR and including interim discussion phases.

Further design experiments with upper secondary students in Austria are planned for 2025. In these, we will evaluate both written responses and recorded oral discussions. It will be particularly

insightful to examine the extent to which students' choices of statistical measures align with and substantiate their fairness approaches, and whether students develop a more critical perspective on the social and ethical consequences of algorithmic decision-making.

REFERENCES

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press. https://fairmlbook.org/

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

Gal, I. (2002). Adults' statistical literacy: Meaning, components, responsibilities. *International Statistical Review*, *70*(1), 1–25. https://doi.org/10.1111/j.1751-5823.2002.tb00336.x

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

Gould, R. (2010). Statistics and the modern student. *International Statistical Review*, *78*(2), 297–315. https://doi.org/10.1111/j.1751-5823.2010.00117.x

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Proceedings of the 30th International Conference on Neural Information Processing Systems* (pp. 3323–3331). Curran Associates Inc. https://dl.acm.org/doi/10.5555/3157382.3157469

Hilger, S., & Büscher, C. (in press). Richness of reflective processes on algorithmic decision-making systems by prospective teachers. *Proceedings of CERME 14*.

O'Neil, C. (2016). *Weapons of Math Destruction. How big data increases inequality and threatens democracy*. Penguin Books.

Orwat, C. (2019). *Risks of Discrimination through the Use of Algorithms*. Federal Anti-Discrimination Agency www.antidiskriminierungsstelle.de/SharedDocs/downloads/EN/publikationen/Studie_en_Diskriminierungsrisiken_durch_Verwendung_von_Algorithmen.pdf?__blob=publicationFile&v=2

Pessach, D., & Shmueli, E. (2020). A Review on Fairness in Machine Learning. *ACM Computing Surveys*, *55*(3), Article 51. https://doi.org/10.1145/3494672

Schönbrodt, S., Schneider, S., Podworny, S., & Camminady, T. (2025). A learning activity on fairness in data-driven algorithmic decision-making systems. *Teaching Statistics. 48,* 33–44. https://doi.org/10.1111/test.70016

Wattenberg, M., Viégas, F., & Hardt, M. (n.d.) *Attacking discrimination with smarter machine learning*. https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Weiland, T. (2017). Problematizing statistical literacy: An intersection of critical and statistical literacies. *Educational Studies in Mathematics*, *96*, 33–47. https://doi.org/10.1007/s10649-017-9764-5

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223–248. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x