Frontline Learning Research
An official journal of EARLI

Frontline **L**earning **R**esearch Vol. 14 No. 1 (2026) 38-70
ISSN 2295-3159

Algorithmic justice in education through de-biasing: towards politically actionable evidence that is rooted in identity theory and de-colonial thought

Adrian Grimm[1], Sebastian Gombert[2], Silvio Armbrüster[3], Marcus Kubsch[4], Anneke Steegh[1], Marianela Navarro Camacho[5], Hannah Kolbe[1], Simon Tautz[1], Karoline Petersohn[1], Valentin Holst[1], Onur Karademir[2], Isabell Bohm[6], Knut Neumann[1]

[1] IPN – Leibniz Institute for Science and Mathematics Education
[2] DIPF – Leibniz Institute for Research and Information in Education
[3] BAuA – Federal Institute for Occupational Safety and Health
[4] FU Berlin – Free University of Berlin
[5] UCR – University of Costa Rica
[6] RUB – Ruhr University Bochum

Article received 7 January 2025 / revised 26 February 2026 / accepted 9 March 2026 / available online 14 September 2026

Abstract

As Artificial Intelligence (AI) algorithms are increasingly used in education, research shows that the use of these algorithms is not without cost. Instead, AI algorithms are prone to biases which are discussed a lot in various domains. The core strength of this contribution is to anchor the discussion of biases in the specific domain of physics education and to discuss the biases in front of a description of the domain-specific inequalities along physics identity development of students. The database consists of the written answers of 527 students to around 30 items from a five-week-period of physics classes in a digital learning environment. Two concrete biases of AI algorithms in physics education and possible approaches to identify and reduce these biases are investigated quantitatively. In a critical discussion from a feminist and de-colonial perspective, it is highlighted that the chosen approaches seem to have promising potentials to mitigate negative effects on under-served students´ physics identity development. Besides, relevant limitations lead to conclusions that additional counter-measures are needed in order to break out of the vicious cycle of reproduction of historically grown inequalities in physics education. The domain-specific analysis can serve as orientation for other domains as well in order to tackle the challenges of AI algorithmic bias effectively and efficiently.

Keywords: de-biasing, pluriverse, STEM identity, slicing analysis, fairness-enhancing strategies, political regulation

Corresponding author: Adrian Grimm, IPN Kiel, grimm@leibniz-ipn.de, *DOI: https://doi.org/10.14786/flr.v14i1.1659*



1 Introduction

1.1 Relevance and contribution

As Artificial Intelligence (AI) algorithms are increasingly used in education, research shows that the use of these algorithms is not without cost. A growing number of studies have revealed, for example, that AI algorithms are prone to racist biases (Cheuk, 2021; Erden, 2020). Hence, mitigating such biases – that is, de-biasing – of algorithms has become increasingly important. From early on, researcher communities have acknowledged the risks related to biases and proposed principles for "ethical use" (Slade, 2016), "non-discrimination" (Bergmann et al., 2019), "fairness" (Diakopoulus et al., 2021), "justice" (Floridi et al., 2018, pp. 696–700), or "equitable treatment of all people" (Cerratto Pargman et al., 2021, p. 2). Despite the existence of such principles, in education, algorithms are rarely de-biased in practice (Kitto & Knight, 2019). In fact, a systematic literature review revealed that most scholarship on bias is rather focused on describing biases than developing strategies to address them, highlighting a significant research gap (L. Li et al., 2023). In other words: We know about biases but do not know how to mitigate them. Not knowing how to mitigate biases is problematic because the mere knowledge of existence of biases is not actionable, neither in the practice of algorithm design nor when it comes to policies for algorithm use.

Previous work has identified substantial challenges when it comes to de-biasing algorithms. Lohaus et al. (2020), for example, demonstrated a major influence of the very definition of bias by showcasing how an algorithm de-biased according to one definition can be understood as biased according to another definition; and Erden (2020) found that algorithms can reproduce historically grown inequalities even if the training data contain no information about race – through the mechanism of proxy discrimination. In educational contexts many historically grown inequalities exist. In the European Union only 28 % of the 2018 graduates in engineering, manufacturing, and construction were women. In Germany, where 79 % of the students from academic households start an academic career, only 12 % of the students whose parents of no professional qualification do so (El-Mafaalani, 2021, pp. 66–67). The reproduction of these inequalities does not take place in the arena of student achievement, in which de-biasing is currently performed (Düchs & Ingold, 2018; OECD, 2016). Instead, scholars suggested that the historically grown inequalities are reproduced in the arena of student identity development (Archer et al., 2015; Avraamidou, 2019; Dou et al., 2019). Global South and Black feminist scholars have stressed that the reproduction of historically grown inequalities is a systemic issue and hence requires a systemic perspective beyond the isolated case in order to address it (Collins, 1990; Crenshaw, 1989; Escobar, 2017; Freire, 1970; Mignolo, 2007). We hence recognise the necessity 1) to conceptualize bias in the context of identity theory and 2) to approach de-biasing from a feminist, de-colonial perspective to make them practically and politically actionable.

This paper showcases how to de-bias algorithms in STEM education by i) focusing on the specific biases relevant in the context of historically grown inequalities in STEM identity development and ii) taking a feminist, de-colonial perspective in the de-biasing of algorithms to account for the under-representation of groups subject to historically grown inequalities in training data. In other words, we bring together three often isolated fields of research in one single theoretical approach: learning analytics, STEM identity development, and de-colonial thought. In our analyses, we draw on two different methods: 1) an examination of the extent to which the training data indeed exhibit bias by predicting the positionalities along diversity dimensions based on the training data and 2) an analysis of the effect of training data containing differing shares of sub-populations, for example investigating the effect of greater shares of female students in training on the testing results for both, female and male students. In the discussion, we highlight how conceptualizing bias in the context of identity theory is mandatory to obtain a valid definition of bias in the respective context and how a de-

colonial feminist perspective can help finding the right strategies to mitigate bias in the training process. Precisely these bias definitions and strategies are needed to reach politically relevant and actionable conclusions. With our research we seek to contribute an exemplar of a de-biasing process rooted in a strong theoretical framework and a sound methodological procedure.

1.2 Authors´ positions based on feminist standpoint theory

"Feminist standpoint theory recognizes that all knowledge is situated in the particular embodied experiences of the knower" (Costanza-Chock, 2020, p. 9). Acknowledging the relevance of standpoints, we position ourselves as authors on diversity dimensions in order to make our standpoints explicit and thereby the positionalities of researchers in certain areas transparent – and in order to open up an opportunity to identify a lack of diversity where diversity would be necessary. At the same time, we value the privacy of each author and therefore do not provide detailed statistics for all diversity dimensions. As a research team, we carry many privileges: We are cis-gendered *white*[1] women and men. We all hold European citizenships and are able-bodied. We acknowledge that these positionalities shape, for example, our perspectives on our data and the questions we ask.

2 Theoretical background

All students should have equal access to education; that is, the opportunity to engage in learning, develop competence and, more importantly, a respective identity (OECD, 2024, p. 9). However, education is subject to a broad range of historically grown inequalities, especially in STEM and physics education. In STEM, for example, although students who identify as female show similar levels of competence as students who identify as male, the latter show substantially stronger aspirations to pursue a career in this domain (OECD, 2016); that is, develop a stronger STEM identity (Cass et al., 2011). A growing number of researchers attribute the diverging identity development to STEM education that marginalizes students who identify as female or, more broadly, students from under-represented groups (Hazari et al., 2020).

2.1 Identity development

The concept of identity and its development has gained traction in science education (Hazari et al., 2020) as issues such as the leaky pipeline, i.e. the disproportionate loss of capable women from STEM disciplines, were not solely explained by established constructs such as interest (Hazari et al., 2010). Identity theory represents a more comprehensive frame for understanding why students engage in STEM, how some students are promoted while others are marginalised and hence a means to work towards more equitable STEM education (Carlone & Johnson, 2007). In principle, STEM identity refers to the ways in which a person navigates the meaning of STEM and how the person positions itself with respect to STEM (Çolakoğlu et al., 2023). Identity development is understood as the process of negotiating the multiple identities within a person (Gee, 2000); including disciplinary identities such as a STEM identity, social identities such as gender (i.e., I am identifying as female, physicists are rarely female), or personal identities such as being a loner (i.e., I am a loner, people liking physics are often loners) (Hazari et al., 2010). The process of developing an identity is driven by students' experiences from prior and current experiences in STEM (Calabrese Barton et al., 2013; Shanahan, 2009). In turn, the development of a STEM identity is considered the major pre-requisite for further engagement in STEM. That is, the process of developing a STEM identity can be described as an iterative process, in which engagement in STEM is a prerequisite for the development of an identity in STEM, and the development of a STEM identity drives further engagement.

[1] We set *white* in italics to emphasize it as a privileged position in the structure of racism rather than a skin colour as done and argued by Black German author Tupoka Ogette in her book "exit racism" (Ogette, 2019, p. 14).

Research has repeatedly documented that a substantial number of students are under-served in terms of developing an identity in STEM or, more specifically, physics. Examples include students who identify as female (Ladewig et al., 2020; Traxler et al., 2016), with low socio-economic status (Bachsleitner et al., 2022; El-Mafaalani, 2021), or of colour (Rainey et al., 2018), as well as students located in the intersections of these dimensions (Avraamidou, 2019; Rosa & Moore Mensah, 2016). A growing body of research examined the mechanisms underlying groups of students being under-served in terms of STEM identity development – for an overview see Çolakoğlu (2023). We do not only face these under-servings in various dimensions, under-representations along various dimensions have been shown to reinforce each other: Bodnar and colleagues (2020), for example, found that while girls, in general, had lower science aspiration scores than boys, Black girls scored lowest in that category. Additionally, the already existing historically grown inequalities pose a particular threat for the development of an identity in STEM, leaving affected students with the conclusion that STEM is not for them and a likely decision for a career outside of STEM. A career in STEM, however, is commonly leading to higher (financial) power and (societal) recognition compared to other occupations. Hence, in order to strengthen diversity in STEM and to broaden access to education for all students, it is necessary to find ways to counter historically grown inequalities with their processes of self-reproduction into future inequalities.

2.2 Interlude: de-colonial thought

> *"For the education offered today to be positive and to have creative potential for Kenya's future it must be seen as an essential part of the continuing national liberation process"* (Ngũgĩ wa Thiong'o, 2011, p. 101).

Countering inequalities in education means addressing matters of justice. Using the words of de-colonial author Ngũgĩ wa Thiong'o, education is political, even the choice of the language used in an educational system is political. From a scientific perspective, we need to be explicit about the justice-relevant assumptions behind our scientific analyses. The decisions about which assumption is the best is a political decision. We see the role of science in informing, for a given political objective, how to best reach that objective given the empirical evidence. In this interlude, we shed light on "what or who has shaped" (Bagga-Gupta, 2025, pp. 14–15) our thinking and for which assumptions we seek to provide a contribution to a path of evidence-based decision making. One important remark: We do focus social justice with a perspective deeply inspired by de-colonial and (Black) feminist thought. We do not mean to address all of the relevant dimensions at once – for instance, we focus on gender identity and not on race in this contribution. However, our discussions are deeply informed by de-colonial and (Black) feminist thought, which is why we highlight these inspirations and explain where we believe that our more specific work is generalisable to the greater field of decolonialities and feminisms – and where it is not.

"My decoloniality is not your decoloniality" (S. Chan, 2023). Chan criticised the lack of differentiating and the impreciseness of researchers which is why we introduce our understandings of decoloniality and (Black) feminist thought. We understand coloniality as colonial continuity, with decoloniality being the "the destruction of the coloniality of world power" (Quijano, 2007, p. 177). In other words, decoloniality addresses not only the formal independence, but the entire "matrix of power" that maintains itself in multiple ways (Folayan, 2025). In deconstructing that matrix of power, the objective shall not be homogenization through what has been criticised as modernisation, but instead to strengthen self-determination against discriminatory power structures (Gómez Torres, 2023; Mignolo, 2007). As the member of the Millennial People Bribri Alí García Segura puts it: Educational systems shall complement diverse identities instead of homogenising and thereby replacing identities (García Segura, 2021; Jara Murillo & García Segura, 2022). What is inspired by de-colonial thinkers can from our perspective be transferred to other oppressive power structures such as sexism, racism, and classism and the countering of respective historically grown inequalities in STEM education. Summarising, we seek to dismantle power structures not only formally but completely. When

addressing inequalities, representation is a relevant objective, but representation needs to be complemented by the objective of self-determination and cultural transformation.

Next to de-colonial thought, (Black) feminist thought has contributed many highly relevant aspects that need to be considered. 1) Social justice is not a-historical but needs to include analyses of the past and active unlearning (hooks, 2003, p. 36). 2) It is not enough to stop discriminative processes – active anti-discriminative work is necessary, for example anti-racist work (Costanza-Chock, 2020, p. 62). 3) Matters of justice include the question of justice for whom. Collins' "matrix of oppression" invites us to analyse inequalities between different positionalities along dimensions such as gender identity (1990). 4) The different dimensions such as race, gender, and class produce specific discrimination at their intersections – and hence sexism can only be addressed in its totality when all other forms of discrimination are addressed as well (Crenshaw, 1989; Hooks et al., 2022). 5) Discrimination cannot be understood as the sum of many isolated single cases, but instead its structural components and its reproduction in institutions such as the education system need to be addressed as well (Ogette, 2019, p. 57). 6) Both de-colonial and feminist struggles are inter-connected at a global level, which expresses a need to not only discuss intersections along various dimensions but also to discuss global implications beyond regional boundaries (Gago, 2019). 7) Under-representations shall not be informed by deficit-oriented analyses of those under-represented, but instead highlight the needs for cultural transformations of institutions well beyond representation only (Kayumova & Dou, 2022, pp. 1097–1099).

The concrete perspective on matters of justice that guides this work is one that embraces both, the aforementioned decolonialialities and feminisms, is one coined by Escobar: the "pluriverse", "a world where many worlds fit" (Escobar, 2017). The pluriverse incorporates not only the objective of reducing historically grown inequalities, but can be understood as a "route into questioning and denaturalising the totalising concept of universality" (Perry & Bradley, 2025, p. 602). The pluriverse is especially well suited for our context of STEM educational systems because we are approaching existing inequalities from a STEM identity perspective. While Pinson and colleagues found that inequalities can be reduced by replacing career choice of students by career assignment based on competence (2020), such options need to be neglected from a pluriversal perspective, asking instead: How can STEM education ecosystems be designed in a way that equally invites students from all positionalities? Identity development thereby remains in the ownership of the students. Besides this huge advantage from an identity perspective, the pluriverse assures to not focus too much on a single issue at hand, losing the entire picture out of sight. The broad focus and search for integration allows to discuss issues of effectiveness and efficiency at the same time. More concretely, a single intervention does not always need to address all critical perspectives. Given the complexity of many diversity dimensions and STEM identity development, the pluriverse is well suited to explore to potentials and limitations of a given intervention and then discuss both in the light of the bigger vision of STEM education ecosystems as worlds where many worlds fit. For any given intervention, the assignment of responsibility of who needs to do the transformative work clearly is with the educational systems and not with the already under-served students (Kayumova & Dou, 2022). Informing pathways for STEM education systems as pluriverses understood in precisely that way is the objective of our contribution.

2.3 Learning analytics and identity development

Besides identity development, recent advances in the field of learning analytics (i.e., the intelligent analysis of data related to student learning) promise a way forward to address historically grown inequalities. AI algorithms provide the opportunity to monitor student learning or learning-related constructs, predict to which extent students will meet the overarching educational goals and offer targeted opportunities to support students' further learning (Karademir et al., 2024). More specifically, AI algorithms can provide students with immediate and personalised feedback about their progress at scale (Dennis et al., 2016; Pardo et al., 2019), allowing for all students to engage STEM



practices together and experience recognition for their competence – supporting the development of a STEM identity (Carlone & Johnson, 2007; Hazari et al., 2010, 2020).

However, AI algorithms also pose a number of threats that may reinforce historically grown inequalities. In fact, AI algorithms are not simply just, ethical or even functioning as intended (Uttamchandani & Quick, 2022). Instead, a rapidly growing amount of research shows how AI algorithms can be racist (Dressel & Farid, 2018), biased towards socio-economic status (Fletcher et al., 2021) or reproduce gender stereotypes (Bolukbasi et al., 2016). This behaviour of AI algorithms is particularly concerning when it affects vulnerable groups such as students from historically under-represented groups. If AI algorithms used for assessment through learning analytics function worse for female students than for male students, the female students, who typically are under-served in terms of opportunities to develop a STEM identity anyhow, will experience insufficient recognition and subsequently are less likely to develop a STEM identity; effectively reducing the number of female students in STEM and thus further increasing historically grown inequalities. A range of recently published studies suggest that this effect is in place for a broad spectrum of vulnerable student groups (Bolukbasi et al., 2016; Cheuk, 2021; Costanza-Chock, 2020; Sha et al., 2022; Traag & Waltman, 2022). In other words, the use of AI algorithms for the purposes of learning analytics without a critical reflection is likely to not live up to its potential for a more equitable education and also bears substantial threats to make education less equitable by reinforcing existing discriminatory practices (Uttamchandani & Quick, 2022).

From a pluriversal perspective, it is necessary to harvest the potentials of learning analytics, both in terms of learning in general as well as for the reduction of historically grown inequalities, while at the same time addressing these threats. A pluriversal lens invites us to ask: Given the transformative pathway towards a pluriverse in front of us for STEM education as a whole, which role should the addressing of threats in learning analytics play in order to have the most efficient of all effective solutions? Which measures against threats in AI algorithms are needed, which are inefficient ways that put unnecessarily high burdens in the way towards harvesting the potentials of AI algorithms for learning? For instance, a very labour-intensive measure against a rather small threat can, also from a pluriversal perspective, be considered as such an unnecessarily high burden.

As a basis for a more critical reflection in the use of AI algorithms for learning analytics purposes, we proposed a framework that describes how learning analytics can affect STEM identity development in Figure 1 (Grimm, Steegh, Çolakoğlu, et al., 2023). The framework links STEM identity development to (responsible) learning analytics. At the center of the framework are STEM identity development opportunities. These opportunities must address three dimensions of STEM identity: recognition, performance, and competence (Carlone & Johnson, 2007). Recognition and performance are particularly relevant to under-served students due to two mechanisms of discrimination, vulnerability and iterability (Grimm, Steegh, Çolakoğlu, et al., 2023). Learning analytics poses potentials and threats to STEM identity development opportunities. Navigating both, potentials and threats, is what “responsible learning analytics” are about (Prinsloo & Slade, 2018). One particular threat arises from the use of AI algorithms for learning analytics that exhibit substantial bias; bias that reflects a misinterpretation of student answers leading to students perceiving their competence wrongly and receiving less recognition.

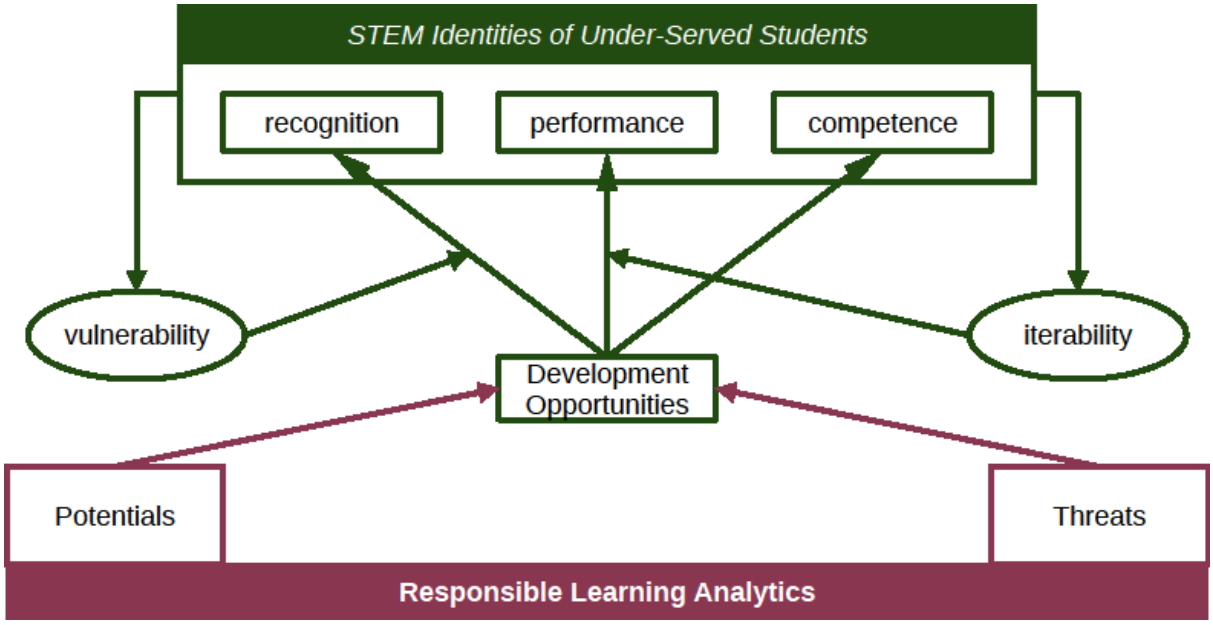


Figure 1 - STEM Identities of Under-Served Students and Responsible Learning Analytics, see (Grimm, Steegh, Çolakoğlu, et al., 2023) for an in-depth explanation of the figure's elements

2.4 Bias in learning analytics

Issues of equity in learning analytics have been approached in terms of the absence of algorithmic bias (Uttamchandani & Quick, 2022). In principle, algorithmic bias refers to situations in which AI algorithms yield a result that advantages or disadvantages specific groups; that is, where decisions made based on the results from the algorithm are discriminatory, violating norms of justice and equity (Kordzadeh & Ghasemaghaei, 2022, p. 388). Algorithmic bias can result, for example, from bias in training data (i.e., specific groups being insufficiently represented in the training data) or the bias(es) of those who develop the algorithms (i.e., specific prejudices, stereotypes or other, often unconscious, attitudes of the developers). Biased algorithms can lead to unjust perceptions, policies, and practices of oppressing under-represented groups (Uttamchandani & Quick, 2022), for example along the dimensions of race (Erden, 2020), ability (Hutchinson et al., 2020) or intersections (Guo & Caliskan, 2021; Tan & Celis, 2019). Computational scientists have developed mathematical techniques to detect and mitigate biases in algorithms and learning analytics researchers have acknowledged the relevance of bias and started to explore bias in AI algorithms used for learning analytics (Cerratto Pargman & McGrath, 2021; Doroudi & Brunskill, 2019; W. Li et al., 2019; Prinsloo & Kaliisa, 2022; Prinsloo & Slade, 2017). Overall, bias in learning analytics algorithms and the threats that come with it are well described (Baker & Hawn, 2021; Erden, 2020; Lohaus et al., 2020; Phillips et al., 2020; Yeung, 2019). However, there is little research to date on how to mitigate bias in learnings analytics.

One reason for a lack of research on how to mitigate bias in learning analytics is likely the very definition of bias. Lohaus and colleagues (2020), for example, demonstrated that an algorithm de-biased according to one definition can be understood as biased according to another definition. Bias has in fact been defined in multiple ways (Baker & Hawn, 2021; Gardner et al., 2019; Mitchell et al., 2021; Suresh & Guttag, 2021). In the context of our aim to achieve a more equitable access and a pluriverse, especially with respect to the development of identity, we define bias as an algorithm yielding "undesirable [...] behaviours or properties" (Cheuk, 2021, p. 2) that lead to a preference for an already privileged group, such as the preference of men through an algorithm, or a discrimination against an already underprivileged group. Note that discrimination against a privileged group (i.e. *white* men) can serve as a justifiable means to achieve greater equity and thus would not be considered a bias according to our definition. This is in line with Constanza-Chock (2020), who argues that for

example racial hierarchies "can only be dismantled by actively anti-racist systems design, not by pretending they don´t exist" (p. 62). Given the historically grown inequalities in STEM education, we see our definition of bias in line with de-colonial and feminist thought as well as the pluriverse. Once such biases are identified, the question arises how these biases can be counteracted; that is by which learning analytics algorithms can be de-biased.

2.5 De-biasing learning analytics

De-biasing learning analytics algorithms requires attending to the different forms of bias that may occur prior to and during the development as well as the use of an algorithm. In Figure 2, which is heavily informed by the work of Baker and Hawn (2021, p. 9), we present an algorithmic life cycle along with entry points of bias; that is, points that de-biasing strategies may be directed at. Before an AI algorithm enters its use phase, many steps need to be taken and in each of these steps, bias can enter the algorithm. The world as is already contains biases, for example historically grown inequalities in the case of physics education. The online world is no perfect representation of the world as it is and the choice of the platform heavily depends on the task an algorithm is set out to perform. Instead of a perfect representation, some parts of the population might not be present on a particular platform. However, an algorithm can only be trained with digital data available. Hence, the data preparation can suffer of under-representation of specific subgroups, the measurement process can introduce biases, and when data is scored as in our context the annotation may introduce biases. The model training itself can introduce new biases when there is no mechanism in place to assure the same performance over all subgroups, for example. Finally, the algorithm may have bias in the use phase in the case of deploying it in a use case for which it was not designed. Two of these biases, namely representation and evaluation bias with the entry points data preparation and model training, bear special relevance to STEM identity development, and hence our work.

Representation bias results from the under-representation of specific sub-groups of the target population in the training data that will be used to train the algorithm. Representation data can originate, for example, from how the data was collected or be an inert part of the data when the sub-groups represent minority groups within the target population (Shahbazi et al., 2023). Since there is fewer information to train the algorithm with, predictions made by the algorithm may be less accurate and hence decisions made based on the predictions can negatively affect students' identity development.

Evaluation bias occurs when the criteria used for evaluating the algorithm differ across sub-groups of the target population. One source of evaluation bias is that the criteria used for evaluation advantage or disadvantage a specific subgroup. In case of evaluation bias, the algorithm may be found to function accurately across all kinds of groups but show high error rates when used with a specific subgroup, with the same effect as representation bias; namely, that students' identity development will be impeded.

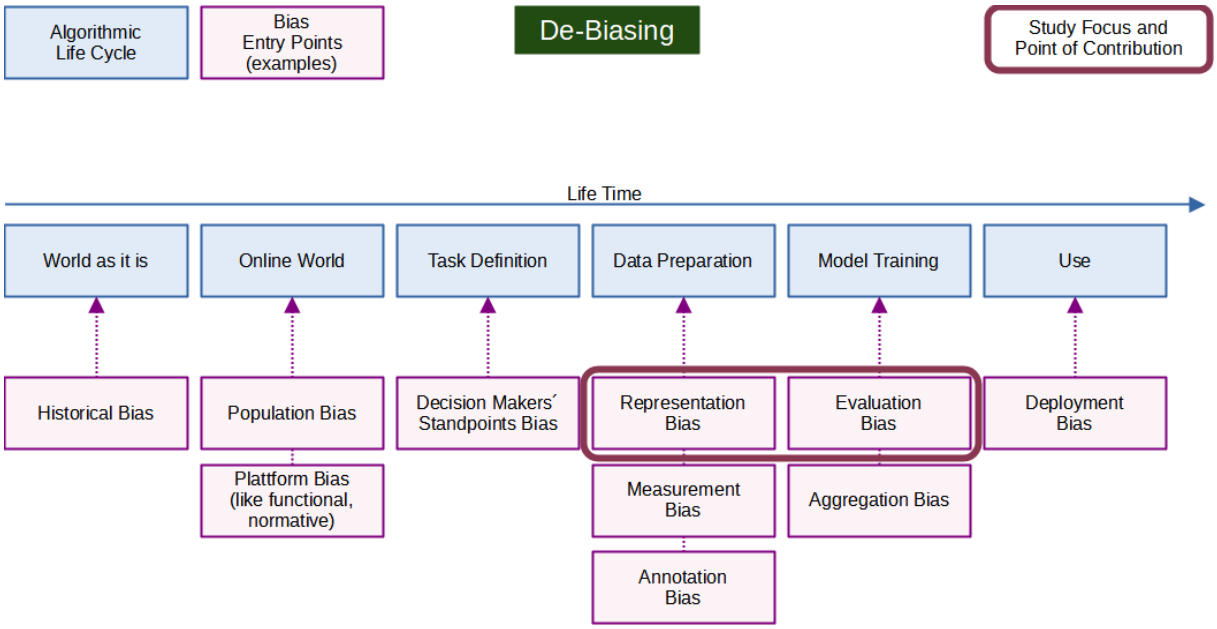


Figure 2 - Bias Entry Points in Learning Analytics and Focus on This Paper (inspired and informed by Baker & Hawn, 2021, p. 9)

Before entering de-biasing strategies, we need to highlight the limits of our research. De-biased algorithms are not bias-free. As we have shown, multiple bias entry points of bias exist and we never address all of them – that is not even the goal. From a pluriversal perspective, the goal is to address the bias entry points wherever they are the most efficient of all effective instruments for reaching de-biased outcomes on a macro level, the STEM identity development of all students. That goal can be translated into: STEM identity development does not depend on a position on any diversity dimension – while students' ownership and self-determination in their identity development are maintained.

In addressing bias, Li and colleagues (2023) identify three groups of de-biasing strategies for machine learning algorithms: 1) pre-processing (i.e. sampling or transforming of the training data to reduce under-representation), 2) in-processing (i.e., designing the algorithm to reduce potential discrimination), and 3) post-processing (i.e., adjusting the prediction outcomes to ensure fairer decisions for those disadvantaged). Pre-processing strategies include for example omitting information about the under-represented group, representing features such that the under-represented group becomes indistinguishable, creating a more balanced training dataset. Creating a more balanced training dataset through oversampling under-represented groups has been found to enhance accuracy of the algorithm for the under-represented group while not negatively affecting accuracy of the algorithm for well represented groups (Sha et al., 2022). In-processing strategies include adding fairness requirements as a regularization term or constraint to the objective function, using an adversary model to minimize the impact of sensitive attributes, or fair representation learning (i.e., acquiring fairer embedding-based representations of the entities involved). Post-processing strategies refer to a broader range of strategies specific to the respective task the algorithm is trying to perform (L. Li et al., 2023, pp. 504–508). Li and colleagues also note from their review of the literature that most work considered (representational) bias in the training data as the main cause for inaccurate decision making. In in our de-biasing analyses, we aim at informing how datasets for algorithmic training need to be set up with respect to diversity dimensions to mitigate representation and evaluation bias. With the objective of reaching a pluriverse in STEM education, we focus on de-biasing learning analytics algorithms in order to strengthen STEM identities especially of under-represented students.



2.6 Research questions

In order to address evaluation bias, we aim at identifying threats of biases before the algorithms are actually trained. In order to do so, we train the same algorithmic architecture with the same student answers but with a different goal, namely to predict positionalities on diversity dimensions, for example to predict students´ gender. The idea is simple: If the algorithm is able to predict gender better, the student answers contain gendered patterns. These patterns could be used by the algorithm in the actual training as well. The greater the gendered patterns, the greater the potential threat of bias. If the algorithm is not able to predict gender, there is no certainty that no gendered patterns exist. However, we expect the probability for bias to be lower. We expect the bias of the actual algorithm used for learning to be higher for the diversity dimensions where the prediction of the positionality on the respective dimension is higher. We ask:

To what extent can threats of bias be identified by training an algorithm to predict positionalities on diversity dimensions?

In order to address representation bias, we seek to find a way to reduce bias based on training dataset configuration when the AI algorithm is trained for its actual task, the prediction of students' learning goals such as knowledge elements like electric energy. From a feminist and de-colonial standpoint, we do not only ask whether the positionalities on diversity dimensions are represented as in the target population. Instead, we aim at finding out how training datasets need to be configured in terms of positionalities on diversity dimensions in order to end up with de-biased algorithms. For example, if we have a lot more students who predominantly speak German instead of another language at home, we do not control for actual target group population shares in our training datasets. Instead, we seek to find out how the representation needs to be configured in the training dataset in order to end up with an algorithm that works at least equally well for students who do not predominantly speak German at home. We ask:

How does dataset slicing effect the prediction results when grouping based on gender, most spoken language at home, or educational background of legal guardians?

3 Methodology

In this work we focus on under-served students, meaning students who face historically grown inequalities and to whom the current physics education systems do not provide enough opportunities to develop their STEM identities. This lack of opportunities provided is what we refer to as under-served in contrast to well-served students with enough opportunities. In the case of our project, a bias in the selected algorithms would lead to incorrect labels for students´ competence. These wrong labels would then be displayed to teacher dashboards and guide interventions in biased directions. A false negative prediction could, for example, result in unwarranted negative feedback from teachers to female and non-binary students possibly leading to a lack of recognition and a weaker STEM identity of these students. A false positive prediction could result in a lack of support that would be necessary for effective learning. Therefore, we decide to use a bias measure that includes correct prediction of both, negative and positive cases.

3.1 Research design

We utilize data from the project "Learning Progression Analytics – Analyzing and Fostering Learning for the Development of Competence" (Grimm et al., 2025). In the project, we implemented one of two so-called curriculum replacement units over five weeks in classes of grade seven or eight

within northern Germany. Both units are implemented in a digital learning environment, have energy transformation as topic, and connect two domains within physics through the energy concept. The units differ in context: One unit is about solar cells with 36 items connecting optics and radiant energy with electricity and electric energy. The other unit is about laptops with 31 items connecting thermodynamics and thermal energy with electricity and electric energy. Each unit follows instruction in line with project-based learning and has one driving question which is split up into three sub-questions. For each sub-question, there is an experiment. The three experimental sessions are framed by an introductory session in the beginning and a summarising session in the end. Before and after the units, we conducted competence tests with the students.

In order to assess students´ competence with our items, we followed evidence-centered design. We formulated a student model with the competence that we want to assess split into sub-dimensions of knowledge, skills, and learning performances. From there, we formulated which evidence we would accept from a student´s answer to an item in order to label the answer with the respective competence element. The important point here is that we ended up by a set of labels on that we scored each student answer – depending on the item there might be one or multiple labels. The students´ answers were then scored by human raters. With the set of students´ answers and scores, we finally trained our algorithms. In Figure 3, you can see an example item from the unit on solar cells. For those who are further interested in the data collection and scoring process, the software scripts or the software versions we used, we provide detailed documentation in the supplemental material.

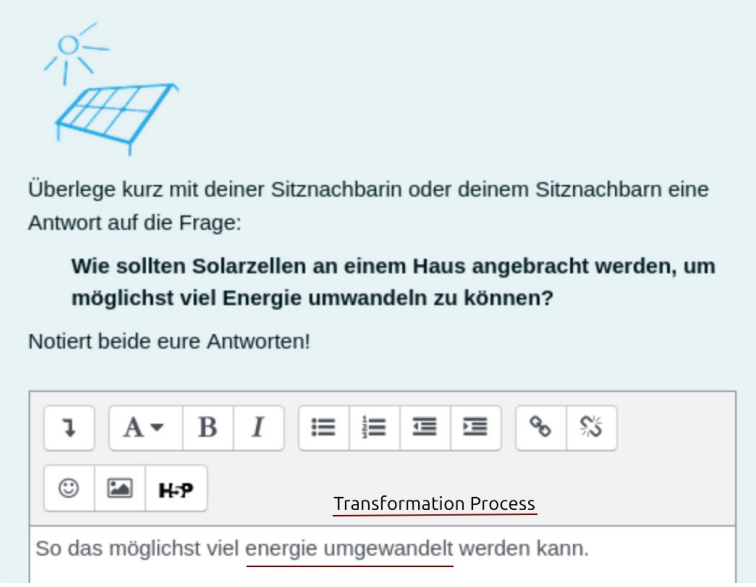


Figure 3 - Example item, answer, label, and score – the item "How should solar cells be installed on a roof in order to transform as much energy as possible?" with the answer "in a way that as much energy can be transformed as possible" with the label "transformation process" scored positively as transformation process identified

We assessed positionalities on aspects of diversity dimensions with items on gender, most spoken language at home, and educational background of legal guardians. These diversity dimensions are not complete, but they are widely used (L. Li et al., 2023) and relevant in the German context (Aikins et al., 2021, p. 69; El-Mafaalani, 2021; Maaz et al., 2022, p. 348) – and therefore suitable to showcase our theoretical and methodological contribution. The assessment was done in the very end of the unit after conducting the post-test in order to not activate stereotype threats. We assessed gender as a diverse item, including options for male, female, non-binary identities, the option to write freely as how the student identifies, and the option to prefer not to answer (gender [gen]). Most spoken language at home is assessed by providing options such as Turkish, German, and a free text field (most spoken language at home [lan]). Educational background of legal guardians is operationalised by whether one of them has an academic degree or not (educational background of legal guardians [edu]).

3.2 Data base

The units were enacted in 22 classes with a total of 527 students. Since some of the teachers we recruited for participation in the original project enacted multiple units in different classes, the 22 classes were located at twelve different schools. The schools were representing the two main tracks in the German school system, which heavily relies on tracking based on academic achievement. Two schools (6 classes) were "Gemeinschaftsschulen" (community schools), representing the lower achieving of the two tracks, 10 schools (16 classes) were Gymnasium schools, representing the higher achieving and university-aiming track.

The student answers were scored by 4 researchers with inter-rater reliabilities´ median of Cohen´s Kappa values for all relevant labels bigger than 0.8. As not all students gave answers to items on diversity dimensions and/or items in the course, the full dataset is reduced for our use case. The number of students who answered the items on each diversity dimension are shown in Table 1 and the respective number of students´ answers that we have for each diversity dimension is shown in Table 2. As you can see, we excluded non-binary genders. We did so as little instances were found on the one hand (which makes analyses impossible) and validity of answers was questionable (at least some students clearly made fun of the category in their answers). Note that this exclusion certainly has implications as evidences for how to support non-binary students are missing – students who are under-served in their STEM identity development. Whether generating big datasets of non-binary students is the most efficient of all effective ways to support them in their STEM identity development needs to be addressed in future works in greater depth. The same holds true for intersectional discrimination (Crenshaw, 1989). Non-binary identities and intersectional discrimination are of high relevance for reaching a pluriverse and, at the same time, needing huge efforts if they shall be addressed in AI system's training. While our contribution does not include empirical evidence on de-biasing for non-binary identities and intersectional discrimination, it can still be understood as a theoretical, methodological, and first empirical step with generalisation potential of indicative character for these discourses within the shared frame of the vision of STEM education pluriverses.

Table 1 - Numbers of Students per Diversity Dimension

diversity dimension	well-served	under-served	total
gender (binary included only)	130	133	263
most spoken language at home	242	40	282
educational background of legal guardians	159	51	210

Table 2 - Numbers of Student Answers per Diversity Dimension

diversity dimension	well-served	under-served	total
gender (binary included only)	1,350	1,391	2,741
most spoken language at home	2,606	381	2,987
educational background of legal guardians	1,724	560	2,284



We scored the student answers on nine different labels in total. Seven labels trace whether a student answer contains a certain knowledge element or not. Two labels trace whether a student answer contains a successful learning performance or not. For further explanations on the labelling process, please refer to the coding book in the supplemental materials.

For our slicing analysis, we train our models label-specific. One model is trained for each label. Additionally, we perform the slicing analyses specific for each of the three dimensions, one dimension for each item. Hence, we end up with 27 slicing analyses resulting from three diversity dimensions times nine labels. In Table 3, we show the available student answers on the level of each of that 27 models. We further distinguish the available positive and negative samples for the labels for each group, well-served and under-served students in the respective diversity dimension.

Table 3 - Numbers of Student Answers per Diversity Dimension Well-Served/Under-Served per Label Positive/Negative; Abbreviations: lan – Language, gen – Gender, edu – Educational Background; ws – Well-Served, us – Under-Served; pos – Positive, neg – Negative; MEE – Manifestation Electric Energy, MEv – Manifestation Electric variable, MRE – Manifestation Radiant Energy, MRv – Manifestation Radiant variable, MTE – Manifestation Thermal Energy, MTv – Manifestation Thermal variable, TP – Transformation Process, M_lp – Manifestation Learning Performance, T_lp – Transformation Learning Performance, colour scheme groups relevant numbers for one model training (27 models in total)

Case	MEE	MEv	MRE	MRv	MTE	MTv	TP	M_lp	T_lp
lan-ws-pos	171	95	319	695	68	199	247	165	192
lan-ws-neg	63	651	820	443	131	50	815	1158	855
lan-us-pos	491	9	51	103	5	6	37	22	30
lan-us-neg	189	87	134	83	4	6	119	171	124
gen-ws-pos	434	47	147	340	25	89	120	74	93
gen-ws-neg	146	328	446	255	68	23	420	607	441
gen-us-pos	411	49	179	367	40	107	141	93	110
gen-us-neg	128	347	432	242	67	29	439	616	461
edu-ws-pos	119	62	234	462	40	132	175	119	143
edu-ws-neg	37	412	491	265	95	36	506	729	525
edu-us-pos	796	22	73	153	17	37	60	46	43
edu-us-neg	265	147	181	100	24	14	174	246	188

3.3 Data analysis procedure

We have one algorithmic architecture with two biases – representation and evaluation bias – to address and three criteria – language, gender, and educational background – that we can use to represent diversity dimensions, ending up with two times three analyses to address bias. In order to

understand the individual datasets, we first provide project context, then describe the full dataset, and finally specify which part of the full dataset we used for the slicing analyses and the training dataset analyses respectively.

3.3.1 Identifying threats of bias: training dataset analysis

In order to address evaluation bias, we implement training dataset analysis. The idea is to train the same model architecture, in our case a transformer model, not on predicting student´s performance but on predicting the diversity dimension. As exemplary shown in Figure 4, we predict the diversity dimension gender instead of predicting the knowledge about electric energy as part of students´ competences. If we find the algorithm to be able to predict gender, we know that the students´ answers to contain what we call gendered patterns. In our case, female students´ answers would contain patterns that distinguish them from male students´ answers.

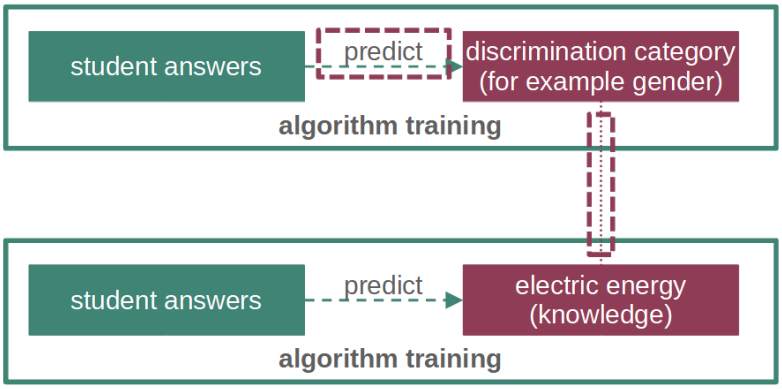


Figure 4 - Identifying Threats of Bias: Training Dataset Analysis

For the training dataset analysis, we use the student answers only and aim at predicting the diversity dimensions. We split the dataset in five splits and then train five models – each split is four times part of the training data and one time forms the testing data. We balance the splits so that each split has as many under- as well-served student answers. For balancing, we use a combined method with over-sampling for the under-served and under-sampling for the well-served student answers. We do so as the combined method for balancing yielded the most reliable predictions on unseen data, thereby preventing over-fitting the best. With our method, we yield four result scores which are Cohen´s kappa, F1-scores[2], precision[3], and recall[4]. We use Cohen´s kappa here instead of quadratic weighted kappa[5] for its more intuitive interpretability – which is more important for a first risk analysis than it is in the more in-depth slicing analysis where quadratic weighted kappa is already well established. We report the arithmetic means for the five models of Cohen´s kappa, F1-score, precision, and recall. The kappa scores allow us to interpret whether a prediction beyond mere guessing is possible and how

[2] F1-scores are a mix of precision and recall: two times precision times recall over the sum of precision and recall.

[3] Precision is an answer to the question: From all the identified positives, how sure can I be that the artefact is actually positive? It is calculated as true positives over the sum of true and false positives.

[4] Recall is an answer to the question: From all actually positive artefacts, how many does my model identify? It is calculated as true positives over the sum of true positives and false negatives.

[5] Quadratic weighted kappa is a score to measure how much prediction outperforms random guessing for a specific dataset and task. It is calculated as difference between empirical probability of agreement and expected agreement with random assignment over the difference between one and the expected agreement with random assignment.

reliably we get that prediction. The F1-score, precision, and recall qualify then how well which student answers are predicted. The interpretation of the scores of all sub-folds remains similar and only varies in effect sizes which is why we do not report them in detail and only mention them here as indicator for reliable results.

3.3.2 Reducing bias: slicing analysis

In order to address representation bias, we implement a slicing analysis inspired by the proposal of Gardner and colleagues and as shown in Figure 5 (2019). The goal is to evaluate how the training datasets need to be set up and who needs to be represented with which share in order to yield satisfactory results. With regard to representation, we focus on these criteria:

- gender
- language that is spoken most at home
- educational background of legal guardians

With "satisfactory results", we refer in our example to de-biased for the specific bias under observation. For all criteria, we split the data in under-served and well-served students – reflecting whether the group faces historically grown inequalities and the system-inherent self-reproduction of these inequalities as discrimination or as privilege. For the example of gender, we check whether training the algorithm only with answers of well-served (here: male) students effects the accuracy for prediction results for under-served (here: female) students. We vary the percentage of under- and well-served students in the training dataset from 0 % to 100 % in 10 %-steps. For slicing, we decide to split on the student level. As one student has more than one answer and training is done on an answer-level, the datasets might vary in size. The training datasets could be further stratified for number of student answers or, as another example, the rating student who could have a decision makers´ standpoint bias, which we do not do in our analyses in order to keep the variations more understandable and interpretable.

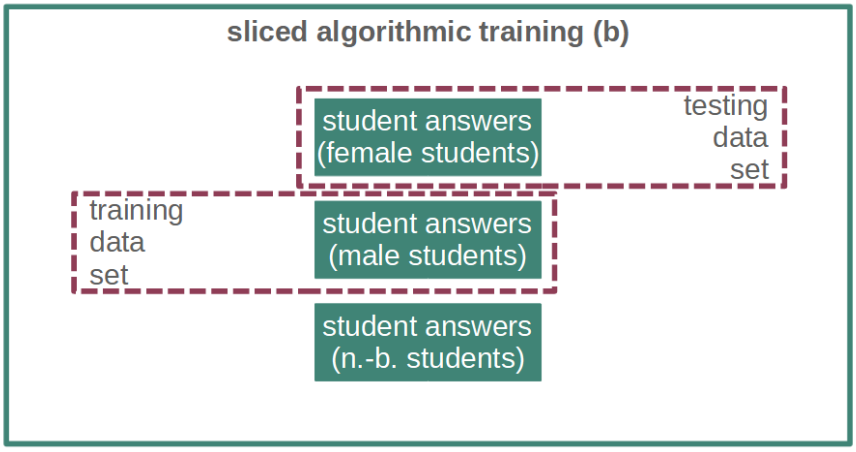


Figure 5 - Reducing Bias: Slicing Analysis

For each of these 27 models, we performed the slicing analyses with eleven variations of the training and testing datasets. We split the datasets eleven times for each diversity dimension with under-served students´ share in the training dataset reaching from 0 % up to 100 % in 10-%-steps. 27 models times eleven splits result in 297 cases for bias evaluation. We stratified label distributions in terms of positive and negative scores for all splits. For each case, we quantified F1-scores, precision, recall, and quadratic weighted kappa for the whole group as well as the sub-groups of well- and under-



served students. For bias evaluation, we refer to bias on F1-scores – using the other scores for further grounding of our interpretations and discussions where needed.

3.4 Algorithmic architecture

In terms of the EU AI Act, we use both general-purpose AI and high-risk systems. In terms of general purpose, we use pre-trained language models and fine-tune these models to our specific use case. For choice of algorithmic architecture, we used the criteria that 1) the models and libraries are frequently used, and 2) the models and libraries are used in many application fields. We built on previous experience in our research project and could therefore use the architecture that we had actually used outside of the context of diversity research (Gombert et al., 2022). We simplified the architecture and lost a bit of cutting-edge technology in order to meet our criterion of frequently used models and libraries. As we are in the application of natural language processing, the models are used across all educational domains and hence are well suited for our second criterion. For those readers who are interested in more details, we used the libraries G-BERT-large (B. Chan et al., 2020) and Huggingface transformers (Wolf et al., 2020).

4 Results

We decided to kick off the results with a very relevant limitation of our findings: We cannot generalise from our evidences alone. Our main contribution is the clear and transparent definition of the process. Our evidences can be first indications and together with many more evidences, they can inform decision making on regulation. However, we provide evidence only for 1) physics education, 2) a specific region of the world, 3) the aspects of diversity dimensions gender, language, and educational background, 4) 7th and 8th grade in school, and 5) the specific topic of energy education – to name some examples. We believe results to be generalisable to a certain extent, as mechanisms can be similar. Nonetheless, this one contribution is far from being enough to draw solid conclusions for political decision making. As our training dataset analysis shall inform the identification of threats of bias, we start with a presentation of our results on the slicing analysis and then present our training dataset analysis, including a discussion in front of the findings from the slicing analysis.

4.1 Reducing bias: slicing analysis

In this sub-section, we present our evidence to the research question:

> *How does dataset slicing effect the prediction results when grouping based on gender, most spoken language at home, or educational background of legal guardians?*

The input of 297 models for training resulted in 132 models that reached a quadratic weighted kappa of bigger than 0.6. Hence, the results presented in the following rely on 132 models only. The quadratic weighted kappa is low for the labels where 1) we have less student answers (for example thermal energy and thermal variables), 2) the label itself holds more content than others (whereas electric energy contains energy constructs only, electric variable contain electric current, voltage, and power – the same holds true for radiant and thermal variable), and 3) the label is of a more complex nature (learning performances). Having small quadratic weighted kappas for these values is thus explainable. Still, the remaining 132 models allow for meaningful slicing analyses in the context of our research question.

4.1.1 Gender

In Figure 6, all 48 bias values for gender are shown with regard to prediction accuracy. We looked at the biases in prediction accuracy as differences in F1 scores between well- and under-served students with positive values representing higher F1 scores for the well-served students.

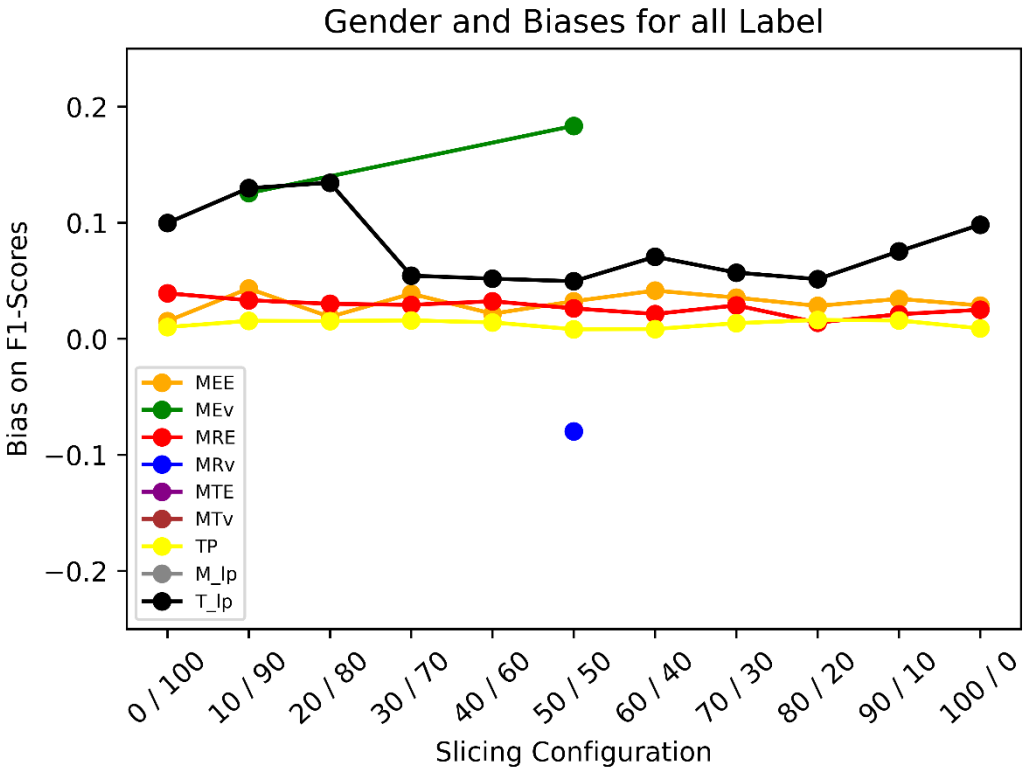


Figure 6 - Gender and Biases for all Label; MEE – Manifestation Electric Energy, MEv – Manifestation Electric variable, MRE – Manifestation Radiant Energy, MRv – Manifestation Radiant variable, MTE – Manifestation Thermal Energy, MTv – Manifestation Thermal variable, TP – Transformation Process, M_lp – Manifestation Learning Performance, T_lp – Transformation Learning Performance

Biases on F1-scores along all slicing configurations for gender reach from -0.080 up to 0.261 with a mean of 0.045 and a median of 0.030. Gender is the only diversity dimension in which we found quadratic weighted kappas bigger than 0.6 for the label of "MEv – Manifestation Electric variable" and "MRv – Manifestation Radiant variable" – though there are still only three values for both labels in total. High kappa values indicate a well-functioning prediction of the respective label and are outside of bias considerations a measure to decide whether a model works or not. Having no values with kappa bigger than 0.6 for the other diversity dimensions means that gender is the only diversity dimension for which we can perform a bias analysis for these labels. Additionally, within the diversity dimension gender also is the only value among all 132 values for biases along all diversity dimensions under observation that is negative – according to our definition of bias and further discriminating under-served students, the only model without bias. In other words: From 297 calculated models, only 132 are considered to work and therefore used for bias analyses. From these 132 models, 131 have a bias: The prediction accuracy for the well-served students is better than the one for the under-served students. No matter what the slicing configuration is, it cannot completely prevent biases in our cases, neither for gender nor for the other diversity dimensions.

There are no clear effects of gender-based dataset slicing on the prediction results in terms of bias on F1-scores. This gender-based evidence supports the claim that balancing training data does not necessarily prevent bias. Differing from Latif and colleagues, our evidence for gender-based slicing does not indicate that balanced datasets outperform imbalanced datasets in terms of bias (2023).

Instead, our evidence on the diversity dimension gender indicates that training dataset configuration is not the most relevant screw to prevent bias. The finding our evidence indicates is: Balancing training datasets does neither prevent nor introduce biases.

However, we find clear indication for existing biases that cannot be explained with the slicing configuration: With a mean of 4.5 % and a median of 3.0 %, our evidence clearly indicates existing biases along the diversity dimension gender. So, we face historically grown inequalities in the world as it is (El-Mafaalani, 2021; Rosa & Moore Mensah, 2016). We are reminded that hooks (2003, p. 36) tells us to consider that history and that Costanza-Chock (2020, p. 62) highlights that we need actively anti-discriminatory interventions. Given that picture of evidences, our evidence indicates a need to further investigate the prevention of bias beyond training dataset configuration so that AI systems do not make worse an already problematic world as it is today.

4.1.2 Most spoken language at home

In Figure 7, all 40 bias values for language are shown.

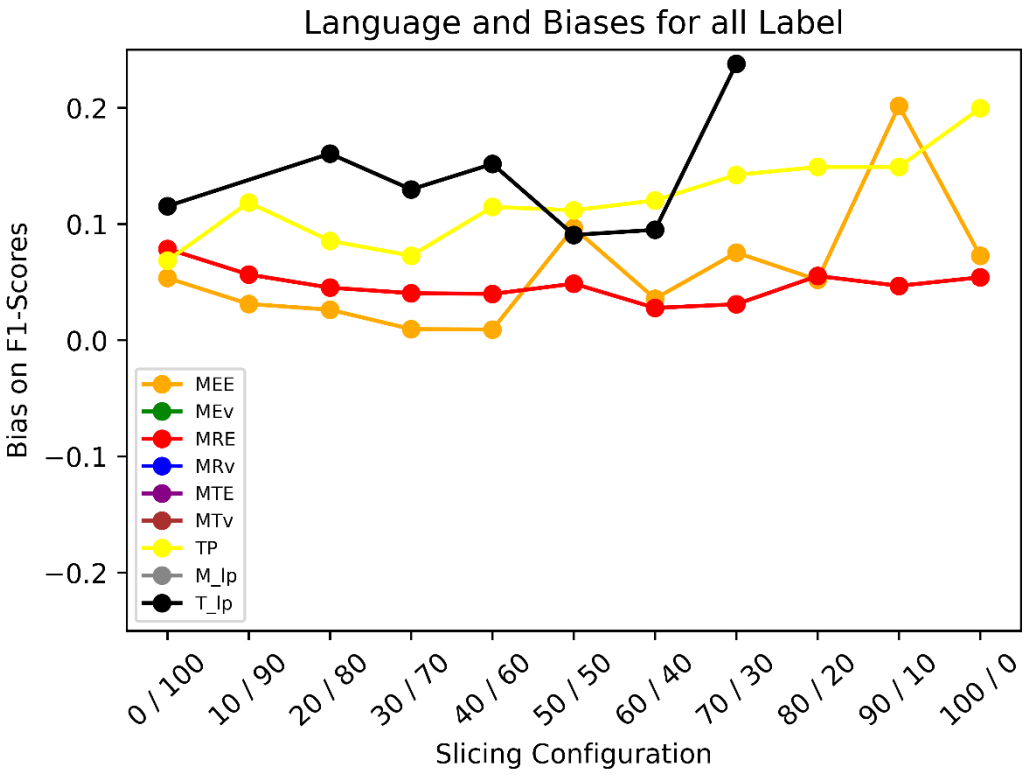


Figure 7 - Language and Biases for all Label; MEE – Manifestation Electric Energy, MEv – Manifestation Electric variable, MRE – Manifestation Radiant Energy, MRv – Manifestation Radiant variable, MTE – Manifestation Thermal Energy, MTv – Manifestation Thermal variable, TP – Transformation Process, M_lp – Manifestation Learning Performance, T_lp – Transformation Learning Performance

Biases on F1-scores along all slicing configurations for most spoken language at home reach from 0.009 up to 0.238 with a mean of 0.087 and a median of 0.074.

Differing from our evidence from gender-based slicing, language-based dataset slicing has an effect on the prediction results in terms of bias on F1-scores. None of the slicing configurations leads to the elimination of bias. Still, slicing configurations yield increasing scores: For example, the average means of the slicing configurations 0 /100 to 20/80, 40 /60 to 60/40, and 80 /20 to 100/0 increase for MEE (0.037, 0.047, 0.109), TP (0.091, 0.116, 0.166), and T_lp (0.135, 0.141, -) while

only those for MRE do not show that trend (0.060, 0.039, 0.052). For MEE as an example, there is a difference in bias from 3.7 % to 10.9 % for the average mean of the respective three slicing configurations. Our evidence here is well in line with the findings from Latif and Colleagues (2023) that balanced datasets outperform imbalanced datasets in terms of bias and that algorithmic biases can be prevented by thoughtful configuration.

The finding our language-based slicing indicates is: Balancing training datasets prevents some biases while not introducing new biases. That finding opens up interesting pathways towards a pluriverse in STEM education, as (at least partly) de-biased AI systems can challenge biased teachers or, more broadly, biased structures (Ogette, 2019, p. 57) in STEM education that prevent under-served students from developing a STEM identity. From the global perspective that Gago (2019) invites, successful legislation in the European Union could also affect de-biasing AI systems in other regions of the world as AI systems are used and traded beyond nation-state boundaries. Hence, even if AI systems cannot be de-biased completely, de-biasing can be one piece within an anti-discrimination architecture that contributes to transform STEM education ecosystems into worlds where many worlds fit. Even more, when de-biasing for languaged patterns could be done without additional data collection efforts by accounting for imbalances in the training process, de-biasing seems to be one of the efficient of all effective solutions.

Similar to our findings on gender-based slicing, we find clear indication for existing biases that cannot be explained with the slicing configuration: With a mean of even 8.7 % and a median of 7.4 % (compared to 4.5 % and 3.0 % for gender-based slicing), our evidence clearly indicates existing biases depending on the most spoken language at home. Our evidence indicates a need to further investigate the prevention of bias beyond training dataset configuration.

4.1.3 Educational background of legal guardians

In Figure 8, all 44 bias values for educational background are shown.

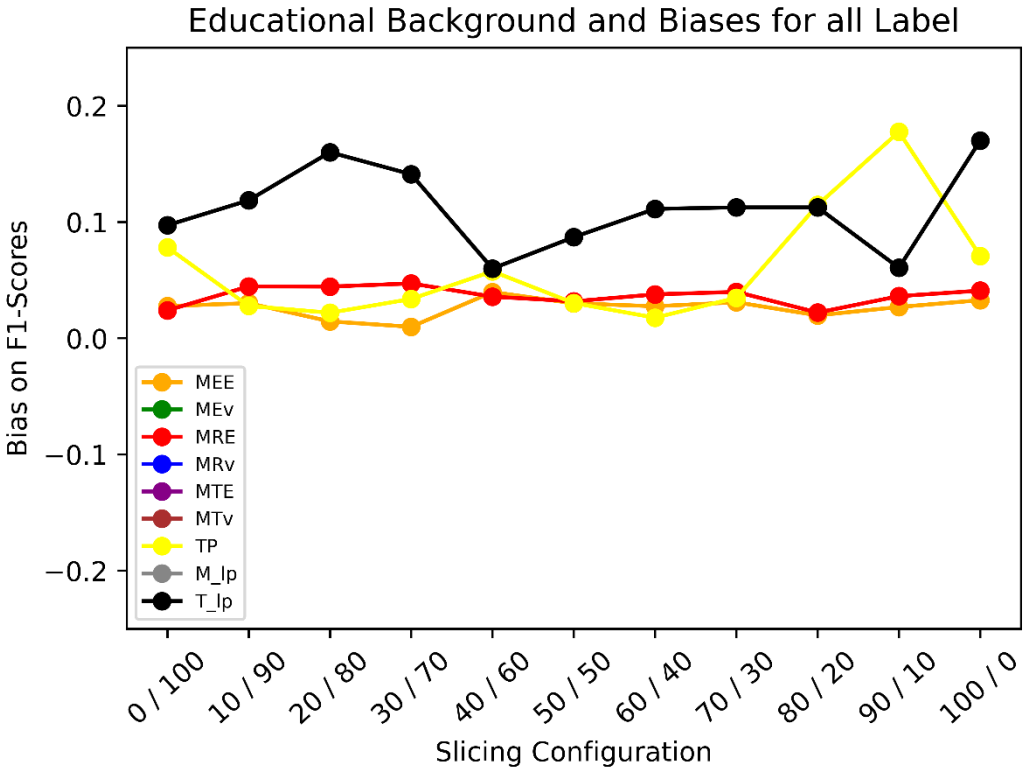


Figure 8 - Educational Background and Biases for all Label; MEE – Manifestation Electric Energy, MEv – Manifestation Electric variable, MRE – Manifestation Radiant Energy, MRv – Manifestation Radiant variable, MTE – Manifestation Thermal Energy, MTv – Manifestation Thermal variable, TP – Transformation Process, M_lp – Manifestation Learning Performance, T_lp – Transformation Learning Performance

Biases on F1-scores along all slicing configurations for educational background of legal guardians reach from 0.010 up to 0.178 with a mean of 0.059 and a median of 0.039.

There are no clear effects of education-based dataset slicing on the prediction results in terms of bias on F1-scores. As for gender-based slicing, our findings indicate that balancing training datasets does neither prevent nor introduce biases.

As for both gender- and language based-slicing, we find clear indication for existing biases that cannot be explained with the slicing configuration: With a mean of 5.9 % and a median of 3.9 %, our evidence clearly indicates existing biases depending on the educational background of legal guardians. Our evidence indicates a need to further investigate the prevention of bias beyond training dataset configuration.

4.2 Identifying threats of bias: training dataset analysis

In this sub-section, we present our evidence to the research question:

To what extent can threats of bias be identified by training an algorithm to predict positionalities on diversity dimensions?

4.2.1 Gender

Table 4 - Results for Gender in Training Dataset Analysis

Score – Arithmetic Means	Under-Served Students	Well-Served Students
kappa	0.126	
F1	0.492	0.588
precision	0.612	0.548
recall	0.456	0.670

In Table 4, the results for gender in our training dataset analysis are shown. The kappa value between 0.0 and 0.2 indicates success beyond mere guessing – even if the prediction is not reliable.

Prediction capability beyond 0.0 tells us that gendered patterns exist in the student answers. Bias is possible. At the same time, gendered patterns in input data do not necessarily lead to a bias in the output. The presented effects also remained in testing with unseen data.

In the light of the slicing analysis where we found biases as well, the need for special care needs to be underlined. In our case, it seems that the algorithm uses the gendered patterns even if trained for label prediction. Hence, the training dataset analysis would have been a valuable risk management measure. However, we need to stress that this case is not generalisable in terms of all high kappa and F1-scores in training dataset analyses need to correlate with biased algorithms. It is well possible that algorithms do not use the gendered patterns and have no bias even though strong gendered patterns exist. Also, a careful dataset configuration would not have led to a de-biased algorithm in our case – other measures need to be found.

4.2.2 Most spoken language at home

Table 5 - Results for Language in Training Dataset Analysis

Score – Arithmetic Means	Under-Served Students	Well-Served Students
kappa	0.789	
F1	0.896	0.892
precision	0.855	0.941
recall	0.943	0.849

In Table 5, the results for language in our training dataset analysis are shown. The kappa value between 0.6 and 0.8 indicates success beyond mere guessing – with remarkably reliable predictions.

Again, prediction capability tells us that language patterns exist in the student answers. Bias is possible but not a necessary consequence. Given the remarkably well predictions, special care for not using precisely these patterns is necessary. However, we need to stress that the remarkably well predictions might be due to over-fitting as the evaluation with unseen data revealed prediction capability that was worse but still existing. The testing with unseen data is an indication and no proof as we tested with very little unseen data (below 100 student answers) which is why we still carefully interpret the results presented in Table 5.

In the light of the slicing analysis where we found slicing effects for language-based slicing, the need for special care needs to be underlined. In our case, it seems that the algorithm uses the languaged patterns even if trained for label prediction. An intervention in terms of careful dataset configuration can de-bias the algorithm. We need to stress that this case is not generalisable in terms of all high kappa and F1-scores in training dataset analyses need to correlate with impact of slicing

analyses. It is well possible that algorithms do not use the languaged patterns and have no bias even though strong languaged patterns exist. What we want to highlight is that in our case, the risk assessment through training dataset analyses could have effectively informed where further slicing analyses are needed. That is precisely the decision that we want to inform: Would a political regulation for training dataset analyses make sense in terms of directing further risk assessment and use of resources? Our evidence can only be seen as a first sign and should not be used to draw conclusions already. Nonetheless, it is a first indication. Whether that indication can be backed by further evidences or not needs to be shown in the future.

4.2.3 Educational background of legal guardians

Table 6 - Results for Educational Background in Training Dataset Analysis

Score – Arithmetic Means	Under-Served Students	Well-Served Students
kappa	0.160	
F1	0.348	0.724
precision	0.627	0.617
recall	0.266	0.886

In Table 6, the results for educational background in our training dataset analysis are shown. The kappa value between 0.0 and 0.2 indicates success beyond mere guessing – even if the prediction is not reliable.

Prediction capability beyond 0.0 tells us that patterns along educational background exist in the student answers. Bias is possible. At the same time, patterns along educational background in input data do not necessarily lead to a bias in the output. We also need to stress that prediction capability vanished when predicting unseen data for educational background which is an indication for over-fitting. It is an indication and no proof as we tested with very little unseen data (below 100 student answers) which is why we still carefully interpret the results presented in Table 6.

Comparing our results of training dataset analyses with our slicing analyses, we find the same pattern as before: Existing biases in the slicing analyses could be traced back in patterns in the student answers along educational background. Little reliability of prediction in the training dataset analyses correlates with little impact of slicing configurations. This evidence is another indication that first risk assessment through training dataset analysis can work.

5 Discussion of implications

In this section, we make our results tangible to three different communities by discussing implications specifically for 1) education researchers, 2) political decision makers, and 3) learning analytics researchers.

5.1 Implications for education researchers

We need education researchers who onboard to discourses around de-biasing algorithmic decision making. Algorithmic decision making comes with great potentials for learning. By no means we aim at downplaying this fact. A good education is a basic right protected in Germany for example through the constitution and at UN level through a Sustainable Development Goal. As community, we need to unpack this potential. At the same time, algorithmic decision making introduces new threats that can reproduce and strengthen historically grown inequalities. However, algorithmic decision making is not inherently biased – how biases can be prevented is an empirical question that can be

answered (Latif et al., 2023). We know from empirical evidence that de-biasing does not necessarily sacrifice predictive accuracy (L. Li et al., 2023, p. 506). Our results indicate that balancing training datasets can be a valuable contribution but alone cannot address all bias issues. Within the education research communities, we build on strong theories to address inequalities, for example identity research in STEM education (Kayumova & Dou, 2022). Combined with the theory on matters of justice, namely the pluriverse with its vision of "a world where many worlds fit", education research communities can meaningfully contribute to the discourses around de-biasing AI systems. These contributions from the education research communities can assure that de-biasing discourses do not fall short in terms of addressing the specific potentials and threats in the context of STEM education.

5.2 Implications for political decision makers

We need political decision makers who aim at setting a regulatory frame for de-biasing algorithmic decision making. We see many funding programmes and grants to unfold the potentials of algorithmic decision making – which we highly appreciate. However, de-biasing practice and the impact of ethical principles have not found their way into practice until today (Kitto & Knight, 2019). In a previous study we have outlined where exactly politically given guidance is missing in order to reach practice: a clear definition of bias, explicit diversity dimensions to analyse for, and a well-defined process of evaluation including evaluation criteria (Grimm, Steegh, Kubsch, et al., 2023). This lack is problematic, as protection against discrimination is an equally strong individual right of each student as the right on education itself. It gets increasingly problematic in domains such as STEM education where already today various historically grown inequalities exist. Positions on diversity dimensions have a strong impact on whether you become a STEM person or not, whether you build a STEM identity or not, and whether you choose a STEM career or not. Algorithmic decision making comes with threats to precisely reproduce these historically grown inequalities. At the same time, there are strong economic interests articulated by powerful, global companies to bring learning analytics to schools. If we do not want to risk to unfold the threats, we need to turn towards concepts such as responsible learning analytics and start building a regulatory frame for de-biasing that addresses the problems where they occur. For example, representation and evaluation bias could be addressed by methods such as slicing and training dataset analyses as proposed in our study. The findings from this study can be first evidence as well as a valuable methodological and theoretical contribution. Nonetheless, much more evidence is needed – for other constructs than energy in physics education, for other domains, for other age groups, and for more diversity dimensions. Political decision makers need to 1) provide funding for and give direction to research to gather further evidence where regulation is most useful and 2) start building up a regulatory frame that brings de-biasing into practice.

5.3 Implications for learning analytics researchers

Learning analytics researchers can focus on bringing de-biasing into practice. The learning analytics from its beginnings has a long and strong history in addressing ethical issues (Prinsloo & Slade, 2017). The most pressing task for the community, from our perspective, is to bring ethics into practical political decision making and to manifest it in a regulatory framework. The focus needs to be on de-biasing strategies and their success in comparison to other strategies. Li and colleagues already reviewed the work on de-biasing which they call "fairness-enhancing strategies", and we need to build on the existing work in the future (2023). It is the task of the community to provide evidences that inform political decision making, that holds true for de-biasing as well. As a community, we need to provide conclusions based on a set of different evaluation criteria and bias definitions that political decision makers can choose from according to their political preferences. We cannot make the political decisions – but we can inform for a given combination of evaluation criteria and bias definition which are the best regulation strategies. That is an empirical question. However, the big global economic players are not going to address them unless they need to and, next to our research communities, not

many actors exist who have the resources to inform political decision making on the regulation of de-biasing. There are questions that need to be answered: How much diversity need the training datasets for successful de-biasing? How many students per position on each diversity dimension do we need for valid and reliable results? Which diversity dimension comes with the biggest risk of bias for a specific domain? Is the regulation of training datasets 1) an effective and 2) the most efficient way to prevent biases? When addressing these questions, we need to carefully reflect and decide how to assess our data, positions on diversity dimensions for instance. Gender can be assessed – as in our example – with more than the binary options. If we assess multiple gender identities, for how many of them do we need to de-bias? As a learning analytics community, we do not need to answer this political question. However, we need to point at the existence of this political question and describe the implications it has. Most important, we need to measure our success as research community not only in terms of theoretical contributions but to closely monitor and steer our impact on practice.

6 Conclusions

6.1 De-biasing through regulation of training datasets – a promising starting point

Regulating training datasets seems to be a promising starting point. Given the few existing evidences, it is too early for final conclusions or recommendations. Nonetheless, existing evidence as well as the evidence in our study support the claim that regulating training datasets can successfully prevent biases in algorithmic decision making. However, our results indicate existing biases that cannot be addressed by the regulation of training datasets alone. Additionally, our findings do not answer the question whether the regulation of training datasets is the most effective way in terms of preventing bias and enabling learning. De-biasing through regulation of training datasets seems to be a promising piece of a regulatory framework that most effectively prevents threats in terms of discrimination and enables both, potentials in terms of learning and diversity main-streaming in physics education.

6.2 Researchers and politicians – everybody can contribute towards a STEM pluriverse

We need 1) education researchers who onboard to discourses around de-biasing algorithmic decision making and contribute strong theoretical frameworks such as the pluriverse and identity development. We need 2) political decision makers who aim at setting an evidence-informed regulatory frame for de-biasing. Finally, we need 3) learning analytics researchers who focus on bringing de-biasing into practice through a) methodological innovation and aiming at actionable results in terms of political decision making, and b) listening for evaluation criteria beyond competence A/B-testing that allow for successfully reaching diversity. Such evaluation criteria can be theoretical contributions formulated by education researchers with identity development and pluriversal perspectives.

Author contributions

Authorship – the earlier the name the greater the contribution for the particular point if no specific contributions are listed: Adrian Grimm (A.G)., Sebastian Gombert (S.G)., Silvio Armbrüster (S.A)., Marcus Kubsch (M.K.), Anneke Steegh (A.S.), Marianela Navarro Camacho (M.N.C.), Hannah Kolbe (H.K.), Simon Tautz (S.T.), Karoline Petersohn (K.P.), Valentin Holst (V.H.), Onur Karademir (O.K.), Isabell Bohm (I.B.), Knut Neumann (K.N.)

Conceptualization: A.G., M.K., A.S., K.N.;

Project operational coordination Work: A.G., O.K., I.B., S.G.;
Data collection: A.G., O.K., I.B., H.K., S.T., K.P.;
Data scoring for algorithmic training: H.K., S.T., K.P., V.H.;
Documentation: H.K., S.T., K.P., A.G., S.A., S.G.;
Discussions about relevant theory: A.G., A.S., M.K., K.N., M.N.C.;
Methodology & results preparation: S.A. (training dataset analyses), S.G. (slicing analyses), A.G. (data pre- and post-processing, further code-commenting);
Original draft preparation: A.G.;
Writing-review and editing: A.G., K.N., A.S., M.K., M.N.C.;
Mentoring: A.S., M.K., M.N.C.;
Project management: A.G., M.K., K.N.;
Funding acquisition: K.N., M.K.;

All authors have read and agreed to the published version of the manuscript.

Acknowledgments

We understand our work as a tiny contribution to the massive house of science built by scientists before us whom we admire, in particular pluriverse thinkers such as Escobar and Mignolo as well as Black feminist thinkers such as hooks and Collins. Also, we want to point out that our work is a quantitatively measurable output that strongly builds on the support of administrative staff at IPN Kiel who are often neither mentioned nor do they receive the merits their work deserves. Finally, we want to highlight our privilege of both having access to so much scientific work and the support of a large administrative department – a too often forgotten privilege, from our perspective.

Funding

The project in which the data for this article were collected received financial support from the Federal Ministry of Education and Research (BMBF), grant number 01JD2008.



References

Aikins, M. A., Bremberger, T., Aikins, J. K., Gyamerah, D., & Yıldırım-Caliman, D. (2021). *Afrozensus 2020: Perspektiven, Anti-Schwarze Rassismuserfahrungen und Engagement Schwarzer, afrikanischer und afrodiasporischer Menschen in Deutschland*. https://afrozensus.de

Archer, L., Dawson, E., DeWitt, J., Seakins, A., & Wong, B. (2015). '"Science Capital"': A Conceptual, Methodological, and Empirical Argument for Extending Bourdieusian Notions of Capital Beyond the Arts. *Journal of Research in Science Teaching*, *52*(7), 992–948. https://doi.org/10.1002/tea.21227

Avraamidou, L. (2019). "I am a young immigrant woman doing physics and on top of that I am Muslim": Identities, intersections, and negotiations. *Journal of Research in Science Teaching*, *57*, 311–341. https:/doi.org/10.1002/tea.21593

Bachsleitner, A., Lämmchen, R., & Maaz, K. (Eds). (2022). *Soziale Ungleichheit des Bildungserwerbs von der Vorschule bis zur Hochschule: Eine Forschungssynthese zwei Jahrzehnte nach PISA*. Waxmann. https://doi.org/10.31244/9783830996248

Bagga-Gupta, S. (2025). Imaginings and Re-imaginings in Doing Multiversal Science. Editor's Introduction. In S. Bagga-Gupta (Ed.), *The Palgrave Handbook of Decolonising the Educational and Language Sciences* (pp. 1–22). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-80322-2_1

Baker, R., & Hawn, A. (2021). *Algorithmic Bias in Education*. https://doi.org/10.1007/s40593-021-00285-9

Bergmann, U., Bonefeld-Dahl, C., Dignum, V., Gagné, J.-F., Metzinger, T., Petit, N., Steinacker, S., Van Wynsberghe, A., & Yeung, K. (Eds). (2019). *Ethics Guidelines for Trustworthy AI*. European Commission - High-Level Expert Group on Artificial Intelligence. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Bodnar, K., Hofkens, T., Wang, M.-T., & Schunn, C. (2020). Science Identity Predicts Science Career Aspiration Across Gender and Race, but Especially for Boys. *International Journal of Gender, Science and Technology*, *12*(1), 32–45.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 4356–4364. https://dl.acm.org/doi/10.5555/3157382.3157584

Calabrese Barton, A., Kang, H., Tan, E., O'Neill, T. B., Bautista-Guerra, J., & Brecklin, C. (2013). Crafting a Future in Science: Tracing Middle School Girls' Identity Work Over Time and Space. *American Educational Research Journal*, *50*(1), 37–75. https://doi.org/10.3102/0002831212458142

Carlone, H. B., & Johnson, A. (2007). Understanding the Science Experiences of Successful Women of Color: Science Identity as an Analytic Lens. *Journal of Research in Science Teaching*, *44*(8), 1187–1218. https://doi.org/10.1002/tea.20237

Cass, C. A. P., Hazari, Z., Cribbs, J., Sadler, P. M., & Sonnert, G. (2011). Examining the impact of mathematics identity on the choice of engineering careers for male and female students. *2011 Frontiers in Education Conference (FIE)*, F2H-1-F2H-5. https://doi.org/10.1109/FIE.2011.6142881

Cerratto Pargman, T., & McGrath, C. (2021). Mapping the Ethics of Learning Analytics in Higher Education: A Systematic Literature Review of Empirical Research. *Journal of Learning Analytics*, *8*(2), 123–139. https://doi.org/10.18608/jla.2021.1

Cerratto Pargman, T., McGrath, C., Viberg, O., Kitto, K., Knight, S., & Ferguson, R. (2021). Responsible Learning Analytics: Creating just, ethical, and caring LA systems. *Companion*

Proceedings. LAK21. https://www.solaresearch.org/wp-content/uploads/2021/04/LAK21_CompanionProceedings.pdf

Chan, B., Schweter, S., & Möller, T. (2020). German's Next Language Model. *Proceedings of the 28th International Conference on Computational Linguistics*, 6788–6796. https://doi.org/10.18653/v1/2020.coling-main.598

Chan, S. (2023). My decoloniality is not your decoloniality: The new multiverse – an opinion piece. *Social Dynamics*, *49*(2), 369–375. https://doi.org/10.1080/02533952.2023.2240151

Cheuk, T. (2021). Can AI be racist? Color-evasiveness in the application of machine learning to science assessments. *Science Education*, 1–12. https:/doi.org/10.1002/sce.21671

Çolakoğlu, J., Steegh, A., & Parchmann, I. (2023). Reimagining informal STEM learning opportunities to foster STEM identity development in underserved learners. *Frontiers in Education*, *8*, 1–16. https://doi.org/10.3389/feduc.2023.1082747

Collins, P. H. (1990). *Black Feminist Thought: Knowledge, Consciousness and the Politics of Empowerment*. https:/doi.org/10.4324/9780203900055

Costanza-Chock, S. (2020). *Design justice: Community-led practices to build the worlds we need*. The MIT Press. https://doi.org/10.7551/mitpress/12255.001.0001

Crenshaw, K. (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*(8), 139–167.

Dennis, M., Masthoff, J., & Mellish, C. (2016). Adapting Progress Feedback and Emotional Support to Learner Personality. *International Journal of Artificial Intelligence in Education*, *26*(3), 877–931. https://doi.org/10.1007/s40593-015-0059-7

Diakopoulus, N., Friedler, S., Arenas, M., Barocas, S., Hay, M., Howe, B., Jagadish, H. V., Unsworth, K., Sahuguet, A., Venkatasubramanian, S., Wilson, C., Yu, C., & Zevenbergen, B. (2021). *Principles for Accountable Algorithms and a Social Impact Statement for Algorithms*. FAT/ML. https://www.fatml.org/resources/principles-for-accountable-algorithms

Doroudi, S., & Brunskill, E. (2019). Fairer but Not Fair Enough On the Equitability of Knowledge Tracing. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 335–339. https://doi.org/10.1145/3303772.3303838

Dou, R., Hazari, Z., Dabney, K., Sonnert, G., & Sadler, P. (2019). Early informal STEM experiences and STEM identity: The importance of talking science. *Science Education*, *103*, 623–637. https://doi.org/10.1002/sce.21499

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580

Düchs, G., & Ingold, G.-L. (2018). Frauenanteil bleibt stabil. *Physik Journal*, *17*(8/9), 32–37.

El-Mafaalani, A. (2021). *Mythos Bildung* (2nd edn). Kiepenheuer & Witsch (KiWi).

Erden, D. (2020). KI und Beschäftigung: Das Ende menschlicher Vorurteile oder der Beginn von Diskriminierung 2.0? In *Wenn KI, dann feministisch* (pp. 77–90). netzforma* eV. https://netzforma.org/publikation-wenn-ki-dann-feministisch-impulse-aus-wissenschaft-und-aktivismus

Escobar, A. (2017). *Designs for the Pluriverse: Radical Interdependence, Autonomy, and the Making of Worlds*. Duke University Press. http://www.jstor.org/stable/j.ctv11smgs6

Fletcher, R. R., Nakeshimana, A., & Olubeko, O. (2021). Addressing Fairness, Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health. *Frontiers in Artificial Intelligence*, *3*, 561802. https://doi.org/10.3389/frai.2020.561802

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds & Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Folayan, D. (2025). From Colonial Legacies to Decolonial Futures: Explorations of Coloniality Through Oxbridge and Lagdan. In S. Bagga-Gupta (Ed.), *The Palgrave Handbook of Decolonising the Educational and Language Sciences* (pp. 665–699). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-80322-2_24



Freire, P. (1970). *Pedagogy of the Oppressed*. Penguin Random House UK.

Gago, V. (2019). *La potencia feminista. O el deseo de cambiarlo todo*. Traficantes de Sueños. https://traficantes.net/sites/default/files/pdfs/TDS_map55_La%20potencia%20feminista_web.pdf

García Segura, A. (2021). *Se' dör stë̈—Somos arte: Las enseñanzas del awá—We are art: The teachings of awá*. International Tree Fund.

Gardner, J., Brooks, C., & Baker, R. (2019). Evaluating the Fairness of Predictive Student Models Through Slicing Analysis. *LAK19: Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 225–234. https://doi.org/https://doi.org/10.1145/3303772.3303791

Gee, J. P. (2000). Identity as an Analytic Lens for Research in Education. *Review of Research in Education*, *25*, 99. https://doi.org/10.2307/1167322

Gombert, S., Di Mitri, D., Karademir, O., Kubsch, M., Kolbe, H., Tautz, S., Grimm, A., Bohm, I., Neumann, K., & Drachsler, H. (2022). Coding energy knowledge in constructed responses with explainable NLP models. *Journal of Computer Assisted Learning*, jcal.12767. https:/doi.org/10.1111/jcal.12767

Gómez Torres, J. (2023). *La educación indígena y las trampas discursivas de la modernidad: Los casos de Guatemala y Costa Rica* (Primera edición). EDUPUC, Editoriales Universitarias Públicas Costarricenses.

Grimm, A., Bohm, I., Karademir, O., Gombert, S., Kubsch, M., Di Mitri, D., Borgards, L., Strauß, S., Drachsler, H., Neumann, K., & Rummel, N. (2025). *Learning Progression Analytics—Analysing and Promoting Learning Progressions to Develop Skills (LPA-AFLEK)* (Version 1.0.0) [Data set]. GESIS. https://doi.org/10.4232/1.14414

Grimm, A., Steegh, A., Çolakoğlu, J., Kubsch, M., & Neumann, K. (2023). Positioning responsible learning analytics in the context of STEM identities of under-served students. *Frontiers in Education*, *7*. https://doi.org/https://doi.org/10.3389/feduc.2022.1082748

Grimm, A., Steegh, A., Kubsch, M., & Neumann, K. (2023). Learning Analytics in Physics Education: Equity- Focused Decision-Making Lacks Guidance! *Journal of Learning Analytics*, *10*(1), 71–84. https://doi.org/10.18608/jla.2023.7793

Guo, W., & Caliskan, A. (2021). Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 122–133. https://doi.org/10.1145/3461702.3462536

Hazari, Z., Chari, D., Potvin, G., & Brewe, E. (2020). The context dependence of physics identity: Examining the role of performance/competence, recognition, interest, and sense of belonging for lower and upper female physics undergraduates. *Journal of Research in Science Teaching*, *57*(10), 1583–1607. https://doi.org/10.1002/tea.21644

Hazari, Z., Sonnert, G., Sadler, P., & Shanahan, M. (2010). Connecting high school physics experiences, outcome expectations, physics identity, and physics career choice: A gender study. *Journal of Research in Science Teaching*, *47*(8), 978–1003. https://doi.org/10.1002/tea.20363

hooks, bell. (2003). *Teaching Community—A Pedagogy of Hope*. Routledge. https://doi.org/10.4324/9780203957769

Hooks, B., Truth, S., Davis, A. Y., The Combahee River Collective, Smith, B., Lorde, A., Hill Collins, P., & Crenshaw, K. (2022). *Schwarzer Feminismus: Grundlagentexte* (N. A. Kelly, Ed.; 2. Aufl). Unrast Verlag.

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020). Social Biases in NLP Models as Barriers for Persons with Disabilities. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5491–5501. https://doi.org/10.18653/v1/2020.acl-main.487

Jara Murillo, C. V., & García Segura, A. (2022). *Sébliwak Francisco García ttö̀ o Las palabras de Francisco García*. Universidad de Costa Rica. https://dibo.site/2024/03/21/sebliwak-francisco-garcia-tto/

Karademir, O., Di Mitri, D., Schneider, J., Jivet, I., Allmang, J., Gombert, S., Kubsch, M., Neumann, K., & Drachsler, H. (2024). I don't have time! But keep me in the loop: Co-designing requirements for a learning analytics cockpit with teachers. *Journal of Computer Assisted Learning*, *40*(6), 2681–2699. https://doi.org/10.1111/jcal.12997

Kayumova, S., & Dou, R. (2022). Equity and justice in science education: Toward a pluriverse of multiple identities and onto-epistemologies. *Science Education*, *106*, 1097–1117. https://doi.org/10.1002/sce.21750

Kitto, K., & Knight, S. (2019). Practical ethics for building learning analytics. *British Journal of Educational Technology*, *50*(6), 2855–2870. https://doi.org/10.1111/bjet.12868

Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388–409. https://doi.org/10.1080/0960085X.2021.1927212

Ladewig, A., Keller, M., & Klusmann, U. (2020). Sense of Belonging as an Important Factor in the Pursuit of Physics: Does It Also Matter for Female Participants of the German Physics Olympiad? *Frontiers in Psychology*, *11*, 2685. https://doi.org/10.3389/fpsyg.2020.548781

Latif, E., Zhai, X., & Liu, L. (2023). AI Gender Bias, Disparities, and Fairness: Does Training Data Matter? *arXiv*. https://doi.org/10.48550/arXiv.2312.10833

Li, L., Sha, L., Li, Y., Raković, M., Rong, J., Joksimovic, S., Neil, S., Gašević, D., & Chen, G. (2023). Moral Machines or Tyranny of the Majority? A Systematic Review on Predictive Bias in Education. *LAK23: 13th International Learning Analytics and Knowledge Conference*, 499–508. https://doi.org/10.1145/3576050.3576119

Li, W., Brooks, C., & Schaub, F. (2019). The Impact of Student Opt-Out on Educational Predictive Models. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 411–420. https://doi.org/10.1145/3303772.3303809

Lohaus, M., Perrot, M., & von Luxburg, U. (2020). Too Relaxed to Be Fair. *Proceedings of Machine Learning Research*, *119*, 6360--6369. https://proceedings.mlr.press/v119/lohaus20a.html

Maaz, K., Artelt, C., Brugger, P., Buchholz, S., Kühne, S., Leerhoff, H., Rauschenbach, T., Schrader, J., & Seeber, S. (2022). *Bildung in Deutschland 2022*. Autor:innengruppe Bildungs berichterstattung. https://doi.org/10.3278/6001820hw

Mignolo, W. D. (2007). Delinking. The rhetoric of modernity, the logic of coloniality and the grammar of de-coloniality. *Taylor & Francis Online*, *21*(2–3), 449–514. https://doi.org/10.1080/09502380601162647

Mitchell, S., Potash, E., D´Amour, A., & Lum, K. (2021). Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, *8*, 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Ngũgĩ wa Thiong'o. (2011). *Decolonising the mind: The politics of language in African literature* (Reprinted). Currey [u.a.].

OECD. (2016). *Excellence and equity in education* (Volume I; PISA 2015 Results). OECD.

OECD. (2024). *Education at a Glance 2024: OECD Indicators*. OECD Publishing. https://doi.org/10.1787/c00cad36-en

Ogette, T. (2019). *Exit racism* (5th edn). unrast-Verlag.

Pardo, A., Jovanovic, J., Dawson, S., Gašević, D., & Mirriahi, N. (2019). Using learning analytics to scale the provision of personalised feedback. *British Journal of Educational Technology*, *50*(1), 128–138. https://doi.org/10.1111/bjet.12592

Perry, M., & Bradley, L. (2025). Pluriversal Literacies: Perspectives and Practices for Sustainable and Anticolonial Futures. In S. Bagga-Gupta (Ed.), *The Palgrave Handbook of Decolonising the Educational and Language Sciences* (pp. 585–612). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-80322-2_21

Phillips, P. J., Hahn, C. A., Fontana, P. C., Broniatowski, D. A., & Przybocki, M. A. (2020). *Four Principles of Explainable Artificial Intelligence*. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8312-draft

Pinson, H., Feniger, Y., & Barak, Y. (2020). Explaining a reverse gender gap in advanced physics and computer science course-taking: An exploratory case study comparing Hebrew-speaking and

Arabic-speaking high schools in Israel. *Journal of Research in Science Teaching*, *57*(8), 1177–1198. https://doi.org/10.1002/tea.21622

Prinsloo, P., & Kaliisa, R. (2022). Learning Analytics on the African Continent: An Emerging Research Focus and Practice. *Journal of Learning Analytics*, 1–18. https://doi.org/10.18608/jla.2022.7539

Prinsloo, P., & Slade, S. (2017). Chapter 4: Ethics and Learning Analytics: Charting the (Un)Charted. In *Handbook of Learning Analytics* (1st edn, pp. 49–57). SoLAR. https://doi.org/10.18608/hla17.004

Prinsloo, P., & Slade, S. (2018). Mapping responsible learning analytics: A critical proposal. In *Responsible Analytics & Data Mining in Education: Global Perspectives on Quality, Support, and Decision-Making*. Routledge.

Quijano, A. (2007). Coloniality and Modernity/Rationality. *Cultural Studies*, *21*(2–3), 168–178. https://doi.org/10.1080/09502380601164353

Rosa, K., & Moore Mensah, F. (2016). Educational pathways of Black women physicists: Stories of experiencing and overcoming obstacles in life. *Physical Review Physics Education Research*, *12*(2), Article 2. https://doi.org/10.1103/PhysRevPhysEducRes.12.020113

Sha, L., Rakovic, M., Das, A., Gasevic, D., & Chen, G. (2022). Leveraging Class Balancing Techniques to Alleviate Algorithmic Bias for Predictive Tasks in Education. *IEEE Transactions on Learning Technologies*, *15*(4), 481–492. https://doi.org/10.1109/TLT.2022.3196278

Shahbazi, N., Lin, Y., Asudeh, A., & Jagadish, H. V. (2023). Representation Bias in Data: A Survey on Identification and Resolution Techniques. *ACM Computing Surveys*, *55*(13s), 1–39. https://doi.org/10.1145/3588433

Shanahan, M.-C. (2009). Identity in science learning: Exploring the attention given to agency and structure in studies of identity. *Studies in Science Education*, *45*(1), 43–64. https://doi.org/10.1080/03057260802681847

Slade, S. (2016). *The Open University Ethical use of Student Data for Learning Analytics Policy*. The Open University. https://doi.org/10.13140/RG.2.1.1317.4164

Suresh, H., & Guttag, J. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *EAAMO '21: Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9. https://doi.org/10.1145/3465416.3483305

Tan, Y. C., & Celis, L. E. (2019). *Assessing Social and Intersectional Biases in Contextualized Word Representations*. Conference on Neural Information Processing Systems. https://proceedings.neurips.cc/paper_files/paper/2019/file/201d546992726352471cfea6b0df0a48-Paper.pdf

Traag, V. A., & Waltman, L. (2022). Causal foundations of bias, disparity and fairness. *arXiv*. https://doi.org/10.48550/arXiv.2207.13665

Traxler, A. L., Cid, X. C., Blue, J., & Barthelemy, R. (2016). Enriching gender in physics education research: A binary past and a complex future. *Physical Review Physics Education Research*, *12*(020114), 1–15. https://doi.org/10.1103/PhysRevPhysEducRes.12.020114

Uttamchandani, S., & Quick, J. (2022). An Introduction to Fairness, Absence of Bias, and Equity in Learning Analytics. In D. Gašević & A. Merceron, *The Handbook of Learning Analytics* (2nd edn, pp. 205–212). SOLAR. https://doi.org/10.18608/hla22.020

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Yeung, K. (2019). *Responsibility and AI* (DGI(2019)05; Issue DGI(2019)05). Council of Europe. https://rm.coe.int/responsability-and-ai-en/168097d9c5