JOURNAL OF LEARNING ANALYTICS

Volume 11(2), 112–122. https://doi.org/10.18608/jla.2024.8301

Adaptive Interventions Reducing Social Identity Threat to Increase Equity in Higher Distance Education: A Use Case and Ethical Considerations on Algorithmic Fairness

Laura Froehlich[1] and Sebastian Weydner-Volkmann[2]

Abstract

Educational disparities between traditional and non-traditional student groups in higher distance education can potentially be reduced by alleviating social identity threat and strengthening students' sense of belonging in the academic context. We present a use case of how Learning Analytics and Machine Learning can be applied to develop and implement an algorithm to classify students as at-risk of experiencing social identity threat. These students would be presented with an intervention fostering a sense of belonging. We systematically analyze the intervention's intended positive consequences to reduce structural discrimination and increase educational equity, as well as potential risks based on privacy, data protection, and algorithmic fairness considerations. Finally, we provide recommendations for Higher Education Institutions to mitigate risk of bias and unintended consequences during algorithm development and implementation from an ethical perspective.

Notes for Practice

- There are educational disparities between traditional and non-traditional students in higher distance education that can be partly explained by social identity threat
- We present a use case of an algorithmic system to identify students at risk of experiencing social identity threat and to present these students with an intervention to reduce it
- Several ethical issues concerning privacy and potential algorithmic bias must be considered during algorithm development and implementation
- We provide recommendations for Higher Education Institutions considering these ethical issues

Keywords: Algorithmic fairness, higher education, student diversity, social identity threat, non-traditional students

Submitted: 17/11/2023 — **Accepted:** 10/05/2024 — **Published:** 04/07/2024

Corresponding author [1]Email: laura.froehlich@fernuni-hagen.de Address: Research Center CATALPA, FernUniversität in Hagen, Universitätsstr. 27, 58097 Hagen, Germany. ORCID iD: https://orcid.org/0000-0001-8610-2750

[2]Email: sebastian.weydner-volkmann@ruhr-uni-bochum.de Address: Ruhr University Bochum, Faculty of Philosophy and Educational Research, Institute for Philosophy I, Ethics of Digital Methods and Technologies, 44780 Bochum, Germany. ORCID iD: https://orcid.org/0000-0003-3948-4770

1. Introduction

Higher education is affected by two trends. First, students come from increasingly heterogeneous sociodemographic backgrounds and the proportion of non-traditional students (i.e., students from historically underrepresented demographic groups in higher education; e.g., Stoessel et al., 2015) is increasing.[1] Second, accelerated by the COVID-19 pandemic, higher education progressively incorporates digital and distance education formats (e.g., Lockee, 2021). These trends make higher

[1] We use the term "non-traditional" students to convey that these groups were historically underrepresented in face-to-face higher education. In distance education, some non-traditional student groups are in the majority (e.g., students older than 30 years, full-time employed students), while others are still underrepresented compared to the general population (e.g., students with migration background, chronic illness, or disability).

education more accessible for sociodemographically diverse students. However, sociodemographic group memberships are related to educational disparities. For example, students with migration background, chronic illness, or first-generation students from lower socio-economic backgrounds show lower performance and higher dropout compared to traditional student groups. Non-traditional students also have more difficulties identifying with their academic institution or developing a sense of belonging in the academic context than traditional students (e.g., Stoessel et al., 2015; Stürmer et al., 2018).

To increase equity in education, Higher Education Institutions (HEIs) strive to reduce group-related disparities in academic success. Academic success should be determined by individual ability rather than by factors related to group membership that present structural hurdles and hinder individual students from fulfilling their academic potential. HEIs might thus seek to implement interventions to reduce group-related educational disparities. One interventional approach is based on social-psychological research on social identity threat. Concerns about being negatively stereotyped due to one's group membership can hinder non-traditional students from fulfilling their academic potential (Steele et al., 2002). Subsequently, we describe current research on the role of social identity threat in explaining educational disparities. We then depict a use case (not a specific implementation) for an adaptive algorithm to determine which students should receive an intervention to alleviate social identity threat. Even though such algorithms are specifically meant to address inequity, their use may still inadvertently (re)produce ethical issues, such as invasion of privacy (Pardo & Siemens, 2014; Prinsloo et al., 2022) or other discriminatory effects (Cerratto Pargman et al., 2023) that countervail their intended positive effect on educational equity. Hence, we analyze the potential for ethical issues with a focus on algorithmic (un)fairness. Lastly, we derive recommendations for developing HEI-specific algorithms and point out limitations and future research directions.

2. Group-Related Educational Disparities Due to Social Identity Threat

Educational disparities between traditional and non-traditional students can partly be explained by negative stereotypes (unfavourable traits or characteristics ascribed to non-traditional student groups). The stereotype content model (e.g., Cuddy et al., 2007) posits that stereotypes can be described on two basic dimensions: warmth (how friendly an outgroup is perceived) and competence (how capable it is perceived). In the educational context, the stereotype dimension of competence is particularly relevant. Student groups that are underrepresented in education (e.g., students with migration background or disability) are often stereotyped as low on competence (e.g., Froehlich et al., 2016; Krischler et al., 2018). Bick et al. (2022) investigated stereotypes about different student groups in higher distance education. In a sample of 626 distance education students, students with migration background, non-native first language, or chronic illness and disability were perceived as low on competence. Such negative stereotypes can have detrimental consequences for members of negatively stereotyped groups. Research on social identity threat (Steele et al., 2002) has shown that if a group is associated with negative stereotypes, this threatens group members' need for a positive social identity. This research is based on social identity theory (Tajfel & Turner, 1979), which states that a person's identity is comprised of a personal identity (i.e., traits and characteristics differentiating the individual from others) and a social identity (i.e., being a member of a social group). Individuals strive for a positive social identity in that they want to feel a positive association between themselves and their group. However, when the group is stereotypically associated with a negative trait like low competence, this threatens the individual's positive social identity. The resulting fear of being associated with the negative traits ascribed to one's group in the academic context is called social identity threat (e.g., Steele et al., 2002). Consequences of social identity threat for affected individuals are reduced academic performance, reduced sense of belonging, disengagement from the academic domain, as well as lower motivation to form relationships with classmates (e.g., Froehlich et al., 2023; Spencer et al., 2016). From a societal perspective, this in turn may lead to a loss in diversity amongst high-performing students and academia in general, and a further reproduction of structural discrimination (e.g., Schmader & Hall, 2014).

Research on social identity threat and its consequences was predominantly conducted in face-to-face learning contexts. Bick et al. (2024) showed that in a distance learning environment with computer-supported collaborative learning, students experiencing higher social identity threat reported a lower sense of belonging to university and the virtual study group, as well as lower motivation to initiate and maintain peer relationships. Therefore, social identity threat can also have detrimental consequences in virtual learning contexts.

3. Interventions to Alleviate Social Identity Threat

Several intervention approaches to reduce social identity threat and its consequences have been developed (Liu et al., 2021). Some approaches aim at creating learning contexts in which social group memberships and associated stereotypes are less salient and therefore do not unfold their negative consequences for non-traditional students. Other approaches target individual learners and aim at strengthening student resilience against social identity threat. One possibility is to strengthen non-traditional students' sense of belonging to the academic domain and thereby reduce the uncertainty of whether they fit in the academic context when their group is underrepresented and negatively stereotyped in this context (Walton & Cohen,

2007). Belongingness interventions often contain statements from other students explaining that belonging uncertainty experienced at the beginning of their studies usually fades over time and is not dependent on minority group membership. Students then write such a statement themselves (e.g., Kizilcec et al., 2020).

Research investigating the effectiveness of interventions to strengthen student resilience against social identity threat has been mainly conducted in face-to-face or laboratory settings, whereas research in digital learning contexts is scarce. Kizilcec et al. (2020) showed that two brief interventions to increase a sense of belonging and decrease social identity threat were effective in eliminating performance differences between students from developed and less developed countries in MOOCs offered by the universities of Harvard and Stanford. The study demonstrated that interventions can reduce group-based performance disparities in asynchronous, computer-mediated learning contexts, which are common in higher distance education.

4. Intended Consequences of Intervention Implementation in Higher Distance Education

Considering this evidence, a major goal to reduce group-based disparities in higher distance education is to use learning analytics (LA) to adaptively present students at risk of experiencing social identity threat with such interventions. Especially in distance education contexts, where students complete individual or group assignments in virtual learning environments, it is possible to monitor learning and performance via LA. For example, machine learning (ML) algorithms can be used to target students most likely to benefit from these interventions.

Operating such a socio-technical intervention system aims at actively strengthening educational equity and supports HEIs in reducing the structural discrimination effects vis-à-vis non-traditional students. This aim will be achieved if the intervention is effective in alleviating social identity threat. Non-traditional students would then be able to perform closer to their intellectual ability. Hence non-traditional students' performance in distance education would be a purer measure of innate ability. To be precise, we must differentiate between two educational performance gaps observable in HEIs: 1) the gap between true intellectual ability and actual performance (or development over time) for individual students that is attributable to the effect of social identity threat; and 2) the educational performance gap between the groups of non-traditional students and traditional students due to structural discrimination. Targeted interventions are assumed to benefit individual students by reducing or closing the first gap. Thereby, interventions are assumed to also have a positive effect on campus diversity by reducing the second gap. This resonates with recent calls in LA ethics to put equity at the centre of LA data science by looking "first at those at the margins of the education system" and by actively "doing ethics" as part of LA data science (Cerratto Pargman et al., 2023, p. 144).

However, the effects of closing the first gap by reducing individual experiences of social identity threat on overall structural discrimination will be limited. The concept of "structural discrimination" is meant to capture effects that consistently keep saliently underprivileged groups in society at a disadvantage — independent of the intent of current individual or institutional actors (Altman, 2020). Hence, structural discrimination can have historical, often diffuse causes. Social identity threat can be understood as both a cause and an effect of structural discrimination: it is a cause since it is an unintended psychological effect that reproduces unjust and disproportionately disadvantageous outcomes; and it is an effect in that, in itself, it is produced as a result of discriminatory structures independent of the intent of current actors in HEIs. The group-related impact of interventions is, hence, expected to be limited to mitigating the causal effects of social identity threat. After interventions, the group of non-traditional students is unlikely to perform completely equal to the group of traditional students. Still, especially due to the empirically established positive effect on individual students, it is worth developing and implementing such an intervention system. In the following section, we describe a corresponding use case.

5. A Use Case of an Algorithm to Recommend an Intervention to Reduce Social Identity Threat to Non-traditional Students

To outline a use case for adaptively reducing the influence of social identity threat on student performance in higher distance education, we will provide a realistic description of an educational context applicable for regular, non-experimental operation. We base this use case on the educational context at the FernUniversität in Hagen in Germany. With more than 68,000 students and a sociodemographically diverse student body, the FernUniversität is like other large distance universities and is thus an ideal educational context for our considerations. Students from non-traditional backgrounds are well represented in the student body. For example, students are on average 39 years old, most of whom are employed while studying (75%), and 10% of whom have a foreign nationality (FernUniversität in Hagen, 2023a). Moreover, approximately 10% report chronic illness or disability (FernUniversität in Hagen, 2023b). Furthermore, a gap between traditional and non-traditional students' performance has been identified (Stoessel et al., 2015) and non-traditional student groups report higher social identity threat than traditional student groups (Bick et al., 2024).

Consequently, such a higher distance education institution might aim at implementing an adaptive intervention to reduce social identity threat, which in turn should increase non-traditional students' performance. Ideally, the group-related performance gap would be reduced, because learning and performance assessments reflect a purer measure of non-traditional students' true ability. The institution thus aims to present the intervention only to those students whose performance and belonging are likely to be negatively affected by social identity threat. By reducing or removing the influence of social identity threat, educational disparities would be levelled, as sociodemographic group memberships associated with negative ability-related stereotypes are less influential for performance and students' true individual ability is better reflected by their educational outcomes (e.g., participation in individual and group assignments, exam enrollment, grades). This notion of fairness aims at achieving educational equity, meaning that pre-existing gaps are narrowed due to the implementation of an intervention that is only presented to student groups experiencing social identity threat.

In our use case, the higher education institution is assumed to seek the development and implementation of an ML algorithm that classifies which students are at risk of experiencing social identity threat and should therefore be assigned to the intervention condition. The ensuing prediction problem can be described as follows: Based on which group-related and individual characteristics should students receive the intervention? This prediction problem arises since not all students belonging to negatively stereotyped non-traditional student groups are at risk of experiencing social identity threat to similar extents. Students vary in their individual susceptibility to social identity threat, which in turn is associated with further psychological constructs. For instance, students who place high value on the negatively stereotyped academic domain or on their membership in the negatively stereotyped group are more prone to experience social identity threat (Spencer et al., 2016). Therefore, an algorithm assigning students to the intervention based only on demographic information would make considerable assignment errors. Consequently, training data for the algorithm should include assessments of psychological constructs like social identity threat and associated variables (e.g., group identification, domain identification).

A crucial issue regarding the viability of the use case concerns its potentially limited prediction accuracy. In the context of social identity threat interventions, it is still unclear whether state-of-the-art ML can develop algorithms that perform considerably better than alternative non-adaptive intervention recommendations. The dataset of the study reported in Kizilcec et al. (2020) was used as training data for ML algorithms to target behavioural science interventions for individual students. However, the developed algorithms failed to produce better student outcomes than assigning the intervention to all or a random selection of students. Kizilcec and Lee (2021) conclude that currently there are prediction limitations of data-driven models and that stakeholders like HEIs should not place too much confidence in algorithmic systems to reduce educational disparities. We believe that it is worth exploring this further, however, as previous data used for algorithm development did not contain direct measures of social identity threat and related psychological constructs. The inclusion of direct measures such as self-reported perceived social identity threat assessed in a survey might increase the algorithm's prediction accuracy.

Addressing such issues related to the technical implementation are planned as a next step in our research process. Currently, however, we focus on describing the use case and systematically reflecting on potential ethical issues concerning algorithmic fairness. Corresponding recommendations about how to tackle these issues can then be directly considered during the development phase (value sensitive design; Franssen et al., 2023). To identify potential ethical issues, the following section highlights unintended consequences of the ML algorithm described in our use case.

6. Unintended Negative Consequences of the Proposed System

6.1. Privacy Impact and Other Ethical Considerations

Any operation of an ML-based intervention system must be expected to create unintended negative consequences — even when the intended impact is ethically laudable. This section describes aspects immediately relevant to the use case to highlight the need for ethical assessment in the design phase. The next section will focus in more detail on aspects related to bias and discrimination as those may directly countervail the ambitions described in the use case.

Since the use case foresees *regular operation*, students are not to be considered *research participants*. While this may be different for the design phase of the intervention system (including the training phase of the classification algorithm), operational LA applications must undergo ethical reflection beyond standards of responsible research in behavioural science, such as those outlined in the Belmont report (National Commission, 1979, pp. 23193–23194) and enforced through ethics review boards (Willis et al., 2016). For example, it has been established early on that LA's data collection, analysis methods, and adaptive interventions resemble the dynamic and structure of surveillance activities such as tracking and profiling (Prinsloo & Slade, 2017; Slade & Prinsloo, 2013).

This highlights the need to reflect potential privacy issues resulting from more personal data on students being collected and analyzed to support the algorithm's classification function as non-traditional students at risk of experiencing social identity threat or not. As Prinsloo et al. (2022) note, privacy considerations should go beyond complying with legal data protection requirements and making use of privacy-enhancing technologies. Rather, they should be responsive to norms of appropriateness governing the flow of information in the local context of the HEI (contextual integrity; Nissenbaum, 2009).

For example, more personal data on students such as self-report measures on group identification or social identity threat should be collected sparsely and be kept confidential.

However, if social identity threat is to be extrapolated from demographics only, its extent might be over- or underestimated for certain students based on imperfect correlations between demographics, group identification, and social identity threat. This affects both the training data for the algorithm (which could, in principle, be obtained through informed consent) and the data needed for operational implementation, as the algorithm cannot be trained to use sets of data substantially different from the data to be classified. While including psychological constructs in the training data and in the operational context will certainly increase the costs of designing and operating the algorithm, its privacy impact in the operational phase may be minimized by running the trained algorithm in a decentralized fashion on student computers to avoid collecting and storing additional personal, potentially sensitive data on HEI servers. The detailed technical implementation of this is beyond the scope of this article. It should be noted, however, that generating lists of identifiable non-traditional students at risk of experiencing social identity threat creates manifold privacy and discrimination risks that need to be thoroughly assessed and mitigated for the training data, but especially for the operational context, where such risks may be prohibitive.

Furthermore, as Drachsler and Greller (2016, p. 95) note, expecting constant, state-of-the-art data security measures from public institutions like universities may be unrealistic. Hence, one may need to start from the assumption that any data gathered in LA-related databases *will at some point in the future be compromised*. Not collecting more sensitive information on HEI servers but leaving them on student machines may be the best option to minimize risks of misuse and hacking.

As opposed to an experimental research setting (which may also include the collection of training data) in which informed consent is the expected norm and can only be withheld in exceptional cases, an operational setting at an HEI may have different demands. Aside from the transparency and accountability requirements, European data protection regulations, for example, allow for forms of legal basis other than consent, especially when HEIs are public institutions and process data in an official capacity laid down in the law. If the intervention is effective in reducing educational disparities, its effect would be maximized if all target students identified by the algorithm to be at risk of experiencing social identity threat participated in the intervention. Hence, it is an ethical question whether participation in the intervention should be mandatory for the students selected or whether students should be given the choice.

As correlations between individual predispositions for the experience of social identity threat and performance outcomes are imperfect, however, there is the risk of false-positive recommendations to participate in the intervention. This would be problematic in the case of adverse effects of the intervention recommendation for some students. It remains an open question what information is necessary for students to make an informed decision about participating in the intervention. Future research should investigate how LA can be used to enable positive interventions while minimizing adverse impacts (Sclater, 2016).

From a practical perspective, since some forms of co-operation are required on behalf of the students (e.g., regarding self-report measures) for the algorithm to work, and since subsequent interventions require participation by students categorized as at risk of experiencing social identity threat, asking for consent seems to be the most viable option. Conducting the interventions on a non-voluntary basis seems like an ethical and legal no-go and would likely be unproductive by rendering the interventions unpopular and the students resistive. Subjecting students to spending time on self-reporting data for classification and, for some, on the interventions would likely only be reasonable when such measures remain the exception on campus: incentivizing students to take part in a plethora of self-report and intervention activities could amount to accosting them.

While there are some ethics frameworks for LA applications (e.g., Drachsler & Greller, 2016; Sclater, 2015), practical implementations of such approaches in education are scarce and a general gap between research and practice exists (Ladjal et al., 2022). While frameworks have considerable merit, we do not believe that it is possible to use them in a simple "checking the boxes" fashion for operational LA applications, as too many contextual factors should be considered during technology design and implementation. This section has addressed some, but not all ethical issues relevant for a comprehensive ethical assessment. Subsequently, we focus on bias and discrimination issues, as they may directly countervail the intended benefits of the intervention.

6.2. Bias and Fairness Aspects of the Classification Algorithm

To ensure that unintended outcomes of the use case do not countervail the intended positive effects, potential biases and discriminatory effects need to be taken into account as the classification algorithm is subject to the same risks of inadvertently (re-)producing inequities as any algorithmic LA application (Simbeck, 2022). Datasets used for training ML-based decision-making algorithms will, thus, usually reflect historic biases underlying structural discrimination (Müller, 2023). At best, such datasets can be made unbiased for one single kind of issue, which makes the common reuse of datasets for other purposes and contexts prone to inadvertently introducing bias (Müller, 2023). Hence, implementations of the use case should refrain from reusing existing datasets for what Kizilcec and Lee (2021) label the measurement and training phases of LA algorithm

development, which precede the action phase, where the algorithm is applied in an operational context and unintended negative effects become manifest.

As Müller (2023) highlights, however, this bias isn't problematic *per se*: it is not the bias itself that is harmful, but the combination of bias and *individual or organizational reliance* on the correctness of the data in decision-making. This highlights that the classification algorithm proposed in the use case should also be considered within its institutional embedding as a socio-technical intervention system, rather than as an isolated piece of software. Hence, as an alternative way of addressing bias and discrimination in the design phase of the algorithm, one could also try to mitigate potential discriminatory impact by relying less on its correctness (e.g., by planning for alternative ways to identify students at risk of experiencing social identity threat or to allow self-enrollment in an intervention). While, in the following, we focus on the algorithm itself, this should be duly noted for actual implementation.

6.2.1. Risk of False-Negative Classification of At-Risk Students

As established above, the algorithm's use case is narrowing the performance gap between traditional and non-traditional students due to structural discrimination (in the sense that social identity threat is one of its effects). Hence, a potential source of countervailing discriminatory effects of the classification algorithm might be that it inadvertently fails to identify distinct subgroups of non-traditional students and, consequently, deprives them of the intervention's positive effect. Are there subgroups of non-traditional students saliently affected by structural discrimination but not classified as at risk of experiencing social identity threat, but who would profit from the intervention, while other subgroups are correctly classified? This seemingly rephrases the classification problem of the algorithm (distinguishing students at risk from students not at risk of experiencing social identity threat), but it is crucial whether, in the action phase, the classification is reliable, to a similar extent, for different salient student groups. As social identity threat is highest for non-traditional student groups who are usually in the minority (e.g., students with migration background, chronic illness, or disability; Bick et al., 2024), intersectional discrimination is of particular relevance.

Addressing this question during the measurement and training phases entails a more detailed definition of non-traditionality and risk of experiencing social identity threat reflecting the local context of the student population at the HEI in question. In the measurement phase, some of the subgroups and intersectional subgroups are likely reduced to only a few cases in the training data or even fully absent, which is prone to result in bias and discriminatory impact (e.g., Chawla et al., 2002). In the model learning phase, this may lead to the prediction accuracy being lower for certain subgroups of non-traditional students.

Conceptually, one may refer to the idea of "algorithmic fairness," which was introduced to answer the growing awareness of such issues in algorithmic decision-making and data analysis (Binns, 2020, p. 515). Algorithmic fairness can be defined as the absence of bias and discriminatory effects in such a system (Kizilcec & Lee, 2021, p. 177). Since complete fairness is a regulative ideal rather than actually obtainable, fairness should "be measured on a continuous and not binary scale, and there are multiple ways to measure fairness" (Kizilcec & Lee, 2021, p. 176). Hence, the field tries to give an operationalized account of *measuring fairness* in ML systems (Binns, 2018, p. 1). From a statistical point of view, the question raised above would be one of adequate sufficiency: "Sufficiency encodes the belief that algorithmic decisions should carry the same level of significance for all groups" (Kizilcec & Lee, 2021, p. 187). In general, however, dealing with intersectionality of multiple kinds of discrimination as well as currently unidentified vulnerable groups poses problems for such kinds of approaches to algorithmic fairness (Binns, 2020, p. 515).

6.2.2. Risks of Being Classified for the Intervention

Ethical considerations are also necessary for algorithm implementation during the action phase (i.e., the operational context). First, the way the intervention is recommended to at-risk students needs to be carefully considered and evaluated. Stigmatizing responses to model predictions are possible when at-risk students are flagged for the intervention in a way that further contributes to discrimination. If non-traditional students have accumulated experiences of social identity threat and group-based discrimination during their educational biographies, being flagged for the intervention might elicit negative responses. Informing students about social identity threat can in itself already work as an intervention against its consequences (e.g., Johns et al., 2005). However, being transparent about the intervention's aim to reduce social identity threat can make negatively stereotyped group memberships more salient and thus exacerbate their influence in the educational context. Communicating to non-traditional students that they were selected for an intervention as they might be affected by social identity threat might further reduce their sense of belonging to the academic context due to highlighting their minority status (e.g., Walton & Cohen, 2007). To date, research empirically investigating different ways to inform participants about being selected for such interventions is lacking.

6.2.3. Risks of False-Positive Classification

Furthermore, it needs to be assessed if *traditional* students from some salient social groups are incorrectly classified more often as at risk of experiencing social identity threat than others (e.g., female as opposed to male students where gender is

not a factor determining the risk of experiencing social identity threat). This can be captured by the statistical measure of separation, which "encodes the belief that a fair algorithm makes correct and incorrect predictions at similar rates for different groups." (Kizilcec & Lee, 2021, p. 186). Similarly to non-traditional, at-risk students, the negative impact of being invited as a *traditional* student to participate in an intervention may vary depending on how the intervention is recommended: Being classified as at risk of experiencing social identity threat may give reason to traditional students who struggle in the introductory phase of their studies due to factors other than social identity threat to question their belonging and whether other members of the HEI see them as non-traditional. Whether this issue of potential negative impact (and, in case of algorithmic unfairness, its unequal distribution across salient social groups) is a mere theoretical possibility or practically relevant has not been established. However, in case of false classifications of traditional students, we believe that this risk may be sufficiently mitigated by openly communicating that the algorithm may be unreliable in this fashion.

Contrary, as (Binns, 2018, pp. 7–9) notes, for some groups the relevant question regarding algorithmic fairness may not be "correctly" identified as at-risk of social identity threat according to a chosen definition adequate to the local context, but rather one of appropriate representation. Some groups may expect to be classified or not classified as such, irrespective of whether this really reflects proneness to social identity threat or self-image as (non-)traditional. A subgroup may consider it important to identify as a traditional constituent group at an HEI, but still be prone to social identity threat. In such a case, it is unclear whether to give precedence to group identity and forego the potential benefits of recommending interventions, or to signal non-traditionality to group members in favour of those benefits. Here, adapting the communication may avoid such dilemmas (e.g., by avoiding the connection to traditionality and focusing on stereotypes), but the choice of action may ultimately be best left to university policy and self-governance.

The third statistical measure proposed by Kizilcec and Lee (2021, p. 186), independence, is meant to reflect "that an algorithm's decision be independent of group membership." Since, however, the goal of the classification is the *identification* of students at risk of experiencing social identity threat, it would not make sense to compare the outcomes across different salient social groups, as differences in the classification rates are to be expected and the absence of such differences would probably indicate a lack of algorithmic reliability.

6.2.4. Group Fairness, Individual Fairness, and Counterfactual Fairness

As Binns (2020, p. 515) notes, algorithmic fairness research began with "fairness measures based on statistical parity between protected groups (e.g., gender, race) in each outcome class, and hence are classed as group fairness measures." The arguments presented above largely follow this line of thinking. While group fairness (GF) is currently the dominant research approach to algorithmic fairness, it "suffers from a number of problems. A wide variety of group fairness definitions have been proposed, and there has been little agreement about which is the most promising" (Fleisher, 2021, p. 481). Furthermore, group fairness measures use a limited number of coarse-grained, prescribed protected groups. This poses challenges for intersectionality and for currently unidentified vulnerable groups (Binns, 2020, p. 515).

Alternative approaches were proposed that are labelled individual fairness (IF) measures, which suggest "that the intuitive notion of fairness is expressed by the principle similar treatment: similar individuals should be treated similarly" (Dwork et al., 2012; Fleisher, 2021, p. 480). Here, two individuals are alike if their:

Combinations of task-relevant attributes are 'nearby' each other in the defined metric space. It is assumed that the distance metric is somehow defined by the people who set the policy (e.g. college admissions tutors), and possibly with broader societal agreement, and that this can be applied to individuals (e.g. college applicants) … Individual fairness says that for any two individuals, if their distance in task-relevant similarity is sufficiently small, they should receive the same outcome. (Binns, 2020, p. 516)

It is sometimes argued that notions of individual and group fairness are in conflict and that using group fairness measures to influence the measurement and model learning phases may violate conditions of individual fairness. In turn, treating similar individuals similarly may not produce outcomes that satisfy group-level notions of fairness (Kizilcec & Lee, 2021, p. 190). However, both Binns (2018, 2020) and Fleisher (2021) contend that the two paradigms are complementary rather than oppositional. As Binns (2020, p. 524) shows, the main challenge in finding the right measures and choosing the right approach (be it IF or GF) lies in the situative description of the potential biases.

Consequently, group fairness measures seem to be a good fit for the described use case. Since the classification problem of the algorithm is itself group-related, we consider it unlikely that individual fairness measures can capture the "non-traditionality" and the "risk of experiencing social identity threat" based on individual properties that serve as distance metrics and that are independent from group membership.

Kizilcec and Lee (2021, p. 190) also note counterfactual measures to algorithmic fairness, where the decision to recommend a social identity threat intervention is fair if the prediction for a member of a certain negatively stereotyped student group remains unchanged in the counterfactual scenario when the student belongs to another negatively stereotyped student group, keeping all other features constant. Again, it seems unsound to assume the algorithm can succeed in classifying students as at risk of experiencing social identity threat independent of actually being a member of a specific student group (i.e., while passing the counterfactual test). Moreover, due to bias arising from limited available variables and

unbalanced samples in the training data, it might occur that no student with a different group membership has the same individual features. In such a case, the algorithm needs to extrapolate outside of the training data set, which can lead to further prediction inaccuracies. Hence, as opposed to group fairness measures, we deem neither individual nor counterfactual fairness measures to be a good fit for the described use case.

7. Discussion: Limitations and Future Directions

Finally, we discuss issues that should be considered during algorithm development and implementation. A first issue concerns the effectiveness of the proposed intervention. If the intended consequence is to mitigate the performance gap between non-traditional and traditional students in higher distance education, the success of this endeavour naturally depends on the effectiveness of the intervention in mitigating the psychological processes leading to social identity threat. If designed in a theory-driven and context-sensitive way, even short one-shot belongingness interventions have lasting effects by triggering a recursive cycle (e.g., Walton, 2014; Walton & Wilson, 2018). However, previous attempts to scale up such educational interventions have produced heterogeneous results. Brief interventions to increase non-traditional students' sense of belonging were particularly effective when the universities or learning environments provided opportunities to experience belonging (Walton et al., 2023). Moreover, contextual factors created heterogeneity in effect sizes when scaling up behavioural science interventions in MOOCs (Kizilcec et al., 2020). In future research, this heterogeneity must be systematically taken into account (Bryan et al., 2021) and ML algorithms are likely to be useful in this endeavour because they can take complex multivariate associations into account when classifying students.

A related question concerns the extent of the second performance gap (between student *groups* due to structural discrimination) that can be reduced by addressing the first gap (between true ability and actual performance for *individual* students due to social identity threat) with an adaptive intervention. It is still unclear what other forms of structural discrimination beyond social identity threat contribute to group-related disparities in higher distance education and how these forms of discrimination can be addressed when investigating the effectiveness of adaptive interventions. Beyond specifically targeting non-traditional students and increasing their resilience against social identity threat, HEIs can also implement measures to decrease the likelihood that stereotypes are activated (e.g., by deleting stereotypical depictions in study materials and learning environments). HEIs can also implement diversity statements and collaborative learning formats to increase intergroup contact and thus reduce stereotypes and prejudice (e.g., Aronson et al., 1978). Taking contextual factors concerning the learning environment and the situation at the HEI into account will also determine the generalizability of the use case to other contexts. For example, previous research has shown that belongingness interventions reduced achievement gaps in MOOCs. Future research thus needs to investigate to what extent these promising results are generalizable.

In the action phase, it is unclear what information is necessary for students to make an informed decision about participating in the intervention. Even small changes in the presentation and administration of interventions can drastically change the results and sometimes even render the interventions ineffective (e.g., Walton & Yeager, 2020). Future research should investigate how to present the intervention while simultaneously reducing potential adverse effects for students that were classified for intervention participation and ensuring that the intervention still achieves the intended effects (Sclater, 2016). Similarly, risks of adverse effects for traditional students who are incorrectly classified for the intervention might be mitigated by communicating that the algorithm-based decision is not 100% accurate. In turn, such communication might impede student trust in LA and the educational benefit of ML, which might also render the intervention ineffective.

8. Ethical Recommendations for HEIs Regarding Algorithms to Reduce Social Identity Threat

- Assess potential ethical risks beyond bias and discrimination (e.g., privacy, data security, consent procedures).
- Oversample underrepresented student groups to increase prediction accuracy.
- Include surveys measuring psychological constructs in training data.
- Think about ways to reduce reliance on the correctness of the algorithms' classifications as part of the socio-technical implementation.
- Apply group fairness measures to show that the underlying data are not neutral and highlight specific areas where the algorithm may need to compensate; consider what is not measured in the data (e.g., psychological variables associated with social identity threat).
- Pay attention to intersectional forms of structural discrimination and algorithmic bias.

- Include feedback loops to optimize the algorithm's prediction accuracy.
- Evaluate: Does the algorithm perform better than random assignment of students to the intervention or control condition? Does the algorithm perform better in reducing educational disparities than simply presenting all students with the intervention?
- Implement discrimination-aware unit tests at each step so that fairness issues can be identified and addressed in a timely and targeted manner (d'Alessandro et al., 2017).
- Consider reasons for data collection and the way the interventions are recommended to prevent countervailing effects.
- Ensure legitimacy through student acceptance and participatory processes: Some ethical conflicts will remain, and it may not be an ethically innocent option to refrain from developing and applying an algorithmic intervention system.
- Consider if some salient social groups would oppose being labelled (non-)traditional and think about potential conflicts in terms of policy and self-governance rather than in prescriptive normative terms.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors declared no financial support for the research, authorship, and/or publication of this article.

References

Altman, A. (2020). Discrimination. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2020 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2020/entries/discrimination/

Aronson, E., Blaney, N., Stephen, C., Sikes, J., & Snapp, M. (1978). *The jigsaw classroom*. SAGE Publications.

Bick, N., Froehlich, L., Friehs, M.-T., Kotzur, P. F., & Landmann, H. (2022). Social evaluation at a distance: Facets of stereotype content about student groups in higher distance education. *International Review of Social Psychology*, *35*(1), 12. https://doi.org/10.5334/irsp.686

Bick, N., Froehlich, L., Voltmer, J.-B., Raimann, J., Reich-Stiebert, N., Seidel, N., Burchart, M., Martiny, S. E., Nikitin, J., Stürmer, S., & Martin, A. (2024). Virtually isolated: Social identity threat predicts social approach motivation via sense of belonging in computer-supported collaborative learning [Manuscript submitted for publication].

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 23–24 February 2018, New York, NY, USA (pp. 149–159). Proceedings of Machine Learning Research. https://proceedings.mlr.press/v81/binns18a.html

Binns, R. (2020). On the apparent conflict between individual and group fairness. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (FAT* '20), 27–30 January 2020, Barcelona, Spain (pp. 514–524). ACM Press. https://doi.org/10.1145/3351095.3372864

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour*, *5*(8), 980–989. https://doi.org/10.1038/s41562-021-01143-3

Cerratto Pargman, T., McGrath, C., Viberg, O., & Knight, S. (2023). New vistas on responsible learning analytics: A data feminist perspective. *Journal of Learning Analytics*, *10*(1), 133–148. https://doi.org/10.18608/jla.2023.7781

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2007). The BIAS map: Behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology*, *92*(4), 631–648. https://doi.org/10.1037/0022-3514.92.4.631

d'Alessandro, B., O'Neil, C., & LaGatta, T. (2017). Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big Data*, *5*(2), 120–134. https://doi.org/10.1089/big.2016.0048

Drachsler, H., & Greller, W. (2016). Privacy and analytics: It's a DELICATE issue: A checklist for trusted learning analytics. *Proceedings of the 6th International Conference on Learning Analytics and Knowledge* (LAK '16), 25–29 April 2016, Edinburgh, UK (pp. 89–98). ACM Press. https://doi.org/10.1145/2883851.2883893

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (ITCS '12), 8–10 January 2012, Cambridge, MA, USA (pp. 214–226). ACM Press. https://doi.org/10.1145/2090236.2090255

FernUniversität in Hagen. (2023a). *Hochschulstatistik* [University statistics]. https://www.fernuni-hagen.de/uniintern/organisation/statistik/index.shtml

FernUniversität in Hagen. (2023b). *Studieren mit behinderung* [Studying with disabilities]. https://www.fernuni-hagen.de/studium/fernuni_fuer_alle/behinderung.shtml

Fleisher, W. (2021). What's fair about individual fairness? *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '21), May 19–21, 2021, Virtual Event, USA (pp. 480–490). ACM Press. https://doi.org/10.1145/3461702.3462621

Franssen, M., Lokhorst, G.-J., & van de Poel, I. (2023). Philosophy of technology. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Spring 2023 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/technology/

Froehlich, L., Brokjøb, L. G., Nikitin, J., & Martiny, S. E. (2023). Integration or isolation: Social identity threat relates to immigrant students' sense of belonging and social approach motivation in the academic context. *Journal of Social Issues*, *79*(1), 264–290. https://doi.org/10.1111/josi.12548

Froehlich, L., Martiny, S. E., Deaux, K., & Mok, S. Y. (2016). "It's their responsibility, not ours": Stereotypes about competence and causal attributions for immigrants' academic underperformance. *Social Psychology*, *47*(2), 74–86. https://doi.org/10.1027/1864-9335/a000260

Johns, M., Schmader, T., & Martens, A. (2005). Knowing is half the battle: Teaching stereotype threat as a means of improving women's math performance. *Psychological Science*, *16*(3), 175–179. https://doi.org/10.1111/j.0956-7976.2005.00799.x

Kizilcec, R. F., & Lee, H. (2021). *Algorithmic fairness in education*. arXiv. https://doi.org/10.48550/arXiv.2007.05443

Kizilcec, R. F., Reich, J., Yeomans, M., Dann, C., Brunskill, E., Lopez, G., Turkay, S., Williams, J. J., & Tingley, D. (2020). Scaling up behavioral science interventions in online education. *Proceedings of the National Academy of Sciences*, *117*(26), 14900–14905. https://doi.org/10.1073/pnas.1921417117

Krischler, M., Pit-ten Cate, I. M., & Krolak-Schwerdt, S. (2018). Mixed stereotype content and attitudes toward students with special educational needs and their inclusion in regular schools in Luxembourg. *Research in Developmental Disabilities*, *75*, 59–67. https://doi.org/10.1016/j.ridd.2018.02.007

Ladjal, D., Joksimović, S., Rakotoarivelo, T., & Zhan, C. (2022). Technological frameworks on ethical and trustworthy learning analytics. *British Journal of Educational Technology*, *53*(4), 733–736. https://doi.org/10.1111/bjet.13236

Liu, S., Liu, P., Wang, M., & Zhang, B. (2021). Effectiveness of stereotype threat interventions: A meta-analytic review. *Journal of Applied Psychology*, *106*(6), 921–949. https://doi.org/10.1037/apl0000770

Lockee, B. B. (2021). Online education in the post-COVID era. *Nature Electronics*, *4*(1), 5–6. https://doi.org/10.1038/s41928-020-00534-0

Müller, V. C. (2023). Ethics of artificial intelligence and robotics. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2023 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/ethics-ai/

National Commission. (1979). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research.* National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, U.S. Department of Health and Human Services. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html

Nissenbaum, H. (2009). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press. https://doi.org/10.1515/9780804772891

Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology*, *45*(3), 438–450. https://doi.org/10.1111/bjet.12152

Prinsloo, P., & Slade, S. (2017). Ethics and learning analytics: Charting the (un)charted. In C. Lang, G. Siemens, A. Wise, & D. Gašević (Eds.), *The handbook of learning analytics* (pp. 49–57). Society for Learning Analytics Research. https://doi.org/10.18608/hla17.004

Prinsloo, P., Slade, S., & Khalil, M. (2022). The answer is (not only) technological: Considering student data privacy in learning analytics. *British Journal of Educational Technology*, *53*(4), 876–893. https://doi.org/10.1111/bjet.13216

Schmader, T., & Hall, W. M. (2014). Stereotype threat in school and at work: Putting science into practice. *Policy Insights from the Behavioral and Brain Sciences*, *1*(1), 30–37. https://doi.org/10.1177/2372732214548861

Sclater, N. (2015, March 3). *A taxonomy of ethical, legal and logistical issues of learning analytics v1.0*. Jisc Involve. https://analytics.jiscinvolve.org/wp/2015/03/03/a-taxonomy-of-ethical-legal-and-logistical-issues-of-learning-analytics-v1-0/

Sclater, N. (2016). Developing a code of practice for learning analytics. *Journal of Learning Analytics*, *3*(1), 16–42. https://doi.org/10.18608/jla.2016.31.3

Simbeck, K. (2022). Künstliche intelligenz und fairness im bildungskontext [Artificial intelligence and fairness in education]. In D. Verständig, C. Kast, J. Stricker, & A. Nürnberger (Eds.), *Algorithmen und autonomie: Interdisziplinäre perspektiven auf das verhältnis von selbstbestimmung und datenpraktiken* [Algorithms and autonomy: Interdisciplinary perspectives on the relationship between self-determination and data practices] (pp. 91–100). Verlag Barbara Budrich.

Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist*, *57*(10), 1510–1529. https://doi.org/10.1177/0002764213479366

Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype threat. *Annual Review of Psychology*, *67*, 415–437. https://doi.org/10.1146/annurev-psych-073115-103235

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. *Advances in Experimental Social Psychology, 34*, 379–440. https://doi.org/10.1016/S0065-2601(02)80009-0

Stoessel, K., Ihme, T. A., Barbarino, M.-L., Fisseler, B., & Stürmer, S. (2015). Sociodemographic diversity and distance education: Who drops out from academic programs and why? *Research in Higher Education*, *56*(3), 228–246. https://doi.org/10.1007/s11162-014-9343-x

Stürmer, S., Ihme, T. A., Fisseler, B., Sonnenberg, K., & Barbarino, M.-L. (2018). Promises of structured relationship building for higher distance education: Evaluating the effects of a virtual fast-friendship procedure. *Computers & Education*, *124*, 51–61. https://doi.org/10.1016/j.compedu.2018.05.015

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33–47). Brooks/Cole.

Walton, G. M. (2014). The new science of wise psychological interventions. *Current Directions in Psychological Science*, *23*(1), 73–82. https://doi.org/10.1177/0963721413512856

Walton, G. M., & Cohen, G. L. (2007). A question of belonging: Race, social fit, and achievement. *Journal of Personality and Social Psychology*, *92*(1), 82–96. https://doi.org/10.1037/0022-3514.92.1.82

Walton, G. M., Murphy, M. C., Logel, C., Yeager, D. S., Goyer, J. P., Brady, S. T., Emerson, K. T. U., Paunesku, D., Fotuhi, O., Blodorn, A., Boucher, K. L., Carter, E. R., Gopalan, M., Henderson, A., Kroeper, K. M., Murdock-Perriera, L. A., Reeves, S. L., Ablorh, T. T., Ansari, S., … Krol, N. (2023). Where and with whom does a brief social-belonging intervention promote progress in college? *Science*, *380*(6644), 499–505. https://doi.org/10.1126/science.ade4420

Walton, G. M., & Wilson, T. D. (2018). Wise interventions: Psychological remedies for social and personal problems. *Psychological Review*, *125*(5), 617–655. https://doi.org/10.1037/rev0000115

Walton, G. M., & Yeager, D. S. (2020). Seed and soil: Psychological affordances in contexts help to explain where wise interventions succeed or fail. *Current Directions in Psychological Science*, *29*(3), 219–226. https://doi.org/10.1177/0963721420904453

Willis, J. E., III., Slade, S., & Prinsloo, P. (2016). Ethical oversight of student data in learning analytics: A typology derived from a cross-continental, cross-institutional perspective. *Educational Technology Research and Development*, *64*(5), 881–901. https://doi.org/10.1007/s11423-016-9463-4