Learning Analytics und Diskriminierung

Nathalie Rzepka, Katharina Simbeck und Niels Pinkwart

Abstract
Mit der zunehmenden Digitalisierung des Lernens werden immer mehr Daten von Lernenden analysiert. Neben zahlreichen Vorteilen ist dabei nicht zu vergessen, dass Learning Analytics oder adaptive Lernsysteme auch Einflüsse haben können, die nicht für alle Lernenden gleichermaßen fair sind. Der Wunsch, Lernsysteme bezüglich ihres Diskriminierungsrisikos zu evaluieren, bringt allerdings weitere Fragen mit sich, die im aktuellen Forschungskontext noch zu wenig hinterfragt werden. Zum einen ist Fairness kein einheitlich definiertes Maß und es gibt zahlreiche Metriken, mit denen man sich der Fairness auf verschiedenen Wegen annähern kann. Zum anderen wird nach wie vor meist die Diskriminierung in Bezug auf Geschlecht oder Ethnie untersucht. Andere demographische Gruppen, Minderheiten oder Charakteristika, die im Bildungskontext eine Rolle spielen, werden kaum erforscht. Letztlich verortet sich Diskriminierung oftmals im Machine Learning-Modell, obgleich es auch zahlreiche andere Stellen gibt, an denen eine Diskriminierung im Erstellungsprozess stattfinden kann. In diesem Artikel werden daher die aktuellen Forschungslücken beschrieben und Vorschläge erarbeitet, diese zu schließen.

1. Einführung

Seit einigen Jahren und spätestens mit Beginn der Covid-19-Pandemie ist die Digitalisierung in der Bildung ›angekommen‹. Digitale Lernplattformen, Apps und Lernmanagementsysteme wie Moodle etablieren sich in allen Lernbereichen, sei es schulisches Lernen, informelles Lernen oder auf Ebene der Hochschulen. Allein durch die Nutzung der vielen digitalen Plattformen werden automatisch Daten erfasst und gespeichert, die die Lernaktivitäten der Nutzenden beschreiben. Durch die automatische Bereitstellung von Lerndaten wird auch deren Auswertung immer einfacher. Das Speichern, Analysieren und Auswerten der Daten wird unter dem Begriff Learning Analytics (LA) zusammengefasst. LA erweitert das digitale Lernen um das Sammeln, Analysieren und Interpretieren der Lerndaten (1st International Conference on Learning Analytics and Knowledge, 2011). Das Ziel von LA ist es, die Lernprozesse zu ›verstehen‹ und dadurch fähig zu sein, die Lernerfahrung

für den Nutzenden zu verbessern (LAK 11, 2011). Zum Feld LA gehören dabei ganz verschiedene Anwendungen: Nutzende können nach spezifischen Kriterien gruppiert werden, die Aktivitäten in Onlinekursen und das Verhalten auf Plattformen können analysiert werden. Ferner kann der Lernfortschritt erfasst, Kompetenzen ermittelt und Vorhersagen über die Nutzenden getroffen werden (Siemens, 2013). Vorhersagen beziehen sich beispielsweise auf das Risiko von Abbrüchen. Je nach Interesse des Lehrenden können nicht nur Leistungen in Bezug auf das Ergebnis evaluiert werden, sondern beispielsweise auch die soziale Zusammenarbeit an sich. Auch die grafische Darstellung von Aktivitäten und Leistungen ist Teil von LA, beispielsweise in Form von Dashboards (Ebner, 2019). Einen Schritt weiter als LA geht das adaptive Lernen. Adaptive Lernumgebungen sind definiert als »Lernumgebungen, die sich in Echtzeit an die Benutzer und ihren Lernstand anpassen« (Meier, 2019, S. 1). Die Aktivitäten der User werden ›überwacht‹, interpretiert und das System reagiert dementsprechend (Paramythis & Loidl-Reisinger, 2003). Die algorithmische Verarbeitung der Daten soll dann zu einer personalisierten Lernerfahrung führen, die auf den bzw. die Nutzenden der Plattform abgestimmt ist. LA und adaptive Lernumgebungen bieten dadurch viele Vorteile. Wie von Bodily und Verbert (2017) zusammengefasst können solche Systeme Empfehlungen für die nächsten Übungen oder Materialien geben, den Lernerfolg vorhersagen und Anpassungen auf Basis der Vorhersage vornehmen, unerwünschtes Verhalten oder Lernschwierigkeiten ausfindig machen, die Reflexion über den Lernprozess fördern sowie die Gefühle der Lernenden ermitteln und dementsprechend reagieren. Obgleich die meisten Systeme nicht alle Möglichkeiten ausschöpfen und nur Teile implementieren, zeigen bereits viele Studien, wie adaptives Lernen den Lernerfolg erhöhen kann. Genannt seien z.B. Arnold und Pistilli (2012), die in ihrer Studie Erfolgsvorhersagen als Ampelanzeige an die Lernenden sendeten und damit die Studierendenbindung erhöhen konnten. Van Oostendorp et al. (2014) verglichen ein adaptives Serious Game mit einem nicht-adaptiven und stellten eine erhöhte Effizienz bei der adaptiven Variante fest. Hooshyar et al. (2018) konnten in einer Studie zeigen, dass die Lesefähigkeiten in Englisch durch ein adaptives Spiel im Vergleich zu einem nicht-adaptiven Spiel verbessert werden konnten. Zugleich gibt es kritische Stimmen, die in ihren Studien keinen relevanten Unterschied zwischen adaptivem Lernen und Quizfragen erkennen konnten (z.B. Griff & Matter, 2013). Andere Forschende erarbeiten die Herausforderungen, welche bei der Einführung von adaptivem Lernen auftreten können (z.B. Mirata et al., 2020). Dazu gehören sowohl technologische Herausforderungen wie die Robustheit, Flexibilität und Benutzerfreundlichkeit solcher Systeme, als auch die Akzeptanz von Stakeholdern. Darüber hinaus liegen weitere Herausforderungen in den benötigten Kapazitäten der Bereitstellung, Wartung und Nutzung von adaptiven Systemen.

Trotz der vielversprechenden Studien und der großen Potenziale für den Bildungssektor wächst mit zunehmender Verwendung von LA und adaptiven Lernsys-

temen in der Praxis die Besorgnis über deren Fairness. Mit steigender Komplexität der Systeme wird es immer schwieriger für Lehrkräfte sowie Tutoren und Tutorinnen, die Prognosen, Darstellungen und Empfehlungen von LA-Systemen nachzuvollziehen. Auch außerhalb des Bildungssektors ist die Diskriminierung in KI-basierten Systemen schon in zahlreichen Studien nachgewiesen worden (z.B. Klare et al., 2012; Sweeney, 2013). Im LA-Bereich zeigt sich im Rahmen von Untersuchungen, dass Lernplattformen die Gefahr mit sich bringen, bestimmte Usergruppen nicht fair zu behandeln (Penn Center for Learning Analytics, 2022). Dabei fanden z.B. Kizilcec und Lee heraus, dass bei der Implementierung von Vorhersagen über den Lernerfolg ein Bias in Bezug auf Geschlecht und Ethnie zu finden ist (Kizilcec & Lee, 2020). Hu und Rangwala (2020) zeigen ähnliche Ergebnisse bei der Vorhersage von Lernerfolg in Bezug auf das Geschlecht. Auch bei der Vorhersage von College Abschlüssen zeigt sich, dass das Modell von Anderson et al. (2019) in Bezug auf männliche Nutzer schlechter funktionierte. Yu et al. fanden 2020 in ihrem Modell zur Vorhersage von Kurz- und Langzeit-Lernerfolg heraus, dass das Modell Frauen fälschlicherweise einen höheren Erfolg vorhersagte als männlichen Studenten. Dies deckt sich in gewisser Weise mit Befunden von Jeong et al. (2022), die in ihrem Modell eine Tendenz feststellten, dass Schüler:innen *of color* schlechtere Vorhersageergebnisse erhalten als weiße Schüler:innen. In all diesen Beispielen werden ganz unterschiedliche Gruppen diskriminiert und bei der Anwendung in der Praxis würden solche Modelle zu Benachteiligungen führen, z.B. wenn sie im Rahmen von Erfolgsvorhersagen genutzt werden, um über die Zulassung von Studierenden zu entscheiden. Es ist also zu sehen, dass KI-basierte LA-Systeme Fluch und Segen zugleich sein können und es stellt sich die Frage, was die Alternative dazu sein könnte – keine intelligenten Lernsysteme verwenden?

Da wir der Ansicht sind, dass dies keine Option ist, liefert der folgende Artikel eine Übersicht über die Problemfelder und Forschungslücken bei der Bewertung und Evaluation von LA oder adaptiven Lernsystemen in Bezug auf Fairness. Dabei wird die Frage des ›richtigen‹ Fairness-Maßes ebenso diskutiert wie der aktuelle Forschungsschwerpunkt in Bezug auf die demographische Gruppe sowie die Verortung von Diskriminierung im Erstellungsprozess von Lernsystemen (Kapitel 2). In Kapitel 3 werden aus den vorgestellten Problemfeldern sieben Leitlinien abgeleitet, die Forschenden bei der Bewertung und Evaluation von Learning Analytics oder adaptiven Lernsystemen in Bezug auf Fairness helfen können.

2. Problemfelder und Forschungslücken

2.1 Die Begriffe verstehen

Um sich der Frage nach den zugrundeliegenden Problemfeldern und Forschungslücken zu nähern, müssen in einem ersten Schritt die Begriffe geklärt werden. Da der Artikel nicht ohne die Begriffe Fairness und Bias auskommt, werden diese im Kontext für die Bewertung von Lernsystemen eingeführt und verwendet.

Fairness

Es lohnt sich, den Begriff Fairness von Systemen im Gegensatz zu Fairness im soziologischen Kontext zu betrachten. Der Duden definiert Fairness als »anständiges Verhalten; gerechte, ehrliche Haltung andern gegenüber« (Duden, 2022). Und schon damit stößt man bei der Bewertung eines Lernsystems an seine Grenzen, denn eine regelbasierte Software oder ein Machine Learning (ML)-System ›verhält‹ sich nicht und hat auch keine Haltung. Im algorithmischen Kontext wurde Fairness daher etwas anders definiert, nämlich als »die Abwesenheit von Vorurteilen oder Bevorzugung einer Person oder einer Gruppe aufgrund ihrer angeborenen oder erworbenen Eigenschaften. Ein unfairer Algorithmus ist ein Algorithmus, dessen Entscheidungen eine bestimmte Gruppe von Menschen begünstigen« (Übersetzt durch die Autor:innen; Mehrabi et al., 2021, S. 115:2).

Bias

Andere Forschende nutzen bei der Bewertung von Lernsystemen bevorzugt den Begriff Bias. Algorithmischer Bias bedeutet hier: »die Vorhersageleistung eines Modells (wie auch immer definiert) unterscheidet sich ungerechtfertigt zwischen benachteiligten Gruppen entlang sozialer Achsen wie Ethnie, Geschlecht und Klasse« (Übersetzt durch die Autor:innen; Mitchell et al., 2021, S. 142). Hierbei bezieht sich Bias jedoch nicht nur auf die Vorhersageleistung, sondern auch auf deren Folgen, beispielsweise daraus resultierenden Interventionen. Andere Forschende nutzen den Begriff Bias, um die statistische Ungleichheit (beispielsweise verschiedene Modellgüten für demographische Gruppen) zu betiteln und verwenden den Begriff der Fairness, um die sozialen und moralischen Implikationen zu beschreiben (Baker & Hawn, 2021). Bei dieser Begriffsnutzung handelt es sich beispielsweise um ein ML-Modell mit einem Bias, wenn dessen Ergebnisse unterschiedlich gut für zwei demographische Gruppen sind. Werden die Vorhersageergebnisse im LA-System genutzt, um beispielsweise Aufgabenempfehlungen auszusprechen, dann wäre dieses System als unfair zu bewerten.

Diskriminierung

Die Benachteiligung einer oder mehrerer Menschen aufgrund ihres Geschlechts, Alters, ihrer Rasse, Religion oder Ähnlichem wird als Diskriminierung definiert (Scherr, 2020). Allgemein definiert ist Diskriminierung sowohl abwertendes Sprechen als auch eine benachteiligte Behandlung (Scherr et al., 2017).

2.2 Das Maß

Mit der Bewusstmachung des Diskriminierungspotenzials von LA-Systemen entsteht der Wunsch nach Auswertung und Messung der Fairness dieser Systeme und den darin implementierten Modellen. Hier stellt sich zunächst die Frage: Nach welchen Kriterien kann Fairness gemessen werden? Bei der Messung von Bias und daraus folgender algorithmischer Diskriminierung gibt es verschiedene Ansatzpunkte, die dabei zu ganz unterschiedlichen Ergebnissen führen.

So gibt es im US-amerikanischen Gesetz die Unterscheidung zwischen *Disparate Treatment* und *Disparate Impact*. Unter *Disparate Treatment* fallen solche Prozesse, die sensible Attribute als Entscheidungsgrundlage nutzen und die Diskriminierung damit intendiert ist (Title VII of the Civil Rights Act of 1964). Bei *Disparate Impact* sind alle Prozesse zusammengefasst, die zwar neutral aussehen, aber dennoch diskriminierende Auswirkungen haben (Title VII of the Civil Rights Act of 1964). Die meisten beschriebenen Probleme mit Bias und Diskriminierung im LA-Kontext sind in die Kategorie *Disparate Impact* einzuordnen (Kizilcec & Lee, 2020). Ein Beispiel für Disparate Treatment ist die Zulassung von Studierenden aufgrund ihres sozio-ökonomischen Status: Das sensible Attribut des sozio-ökonomischen Status wird als Entscheidungsgrundlage zur Zulassung genutzt. Disparate Impact hingegen wäre es, wenn ein Modell die Abbruchquote der Studierenden prognostizieren und dabei den Wohnort in die Berechnung mit einbeziehen würde. Da es Wohnorte mit höherem oder niedrigerem sozio-ökonomischen Durchschnitt gibt, kann ein solches Modell diskriminierende Auswirkungen haben.

Eine Berechnung, die angewendet werden kann, ist die Überprüfung nach der demographic parity. Diese beschreibt, dass für jede demographische Gruppe gleich viele Vorhersagen oder Klassifizierungen mit demselben Ergebnis getroffen werden sollen (Gardner et al., 2019). Hier stellt sich allerdings die Frage, ob das tatsächlich fair ist: schaut man beispielsweise in den Bereich der Rechtschreibung oder Lesekompetenz, so ist deutlich, dass Mädchen hier durchschnittlich besser sind als Jungen (Schiepe-Tiska et al., 2016; Valtin et al., 2003). Eine Bias-Definition auf Basis von demographic parity würde aber erwarten lassen, dass es gleich viele positive Vorhersagen gibt für Jungen wie für Mädchen. Demographic parity ist also kein Maß das die individuelle Fairness herstellt, denn auch sensitive Attribute können mit einer Zielvariable korrelieren (Dwork et al., 2012; Lipton et al., 2018).

Viele Studien greifen bei der Überprüfung von Diskriminierung in Systemen und Modellen des maschinellen Lernens daher auf Metriken zurück, die auf der Konfusionsmatrix basieren (Riazy & Simbeck, 2019; Seyyed-Kalantari et al., 2021; Verma & Rubin, 2018). Tabelle 1 zeigt die Konfusionsmatrix, die zwei Dimensionen vereint: die Vorhersage, die entweder positiv oder negativ sein kann, und der wahre Wert, der entweder positiv oder negativ sein kann. Ein Wert ist richtig positiv, wenn sowohl der wahre Wert als auch die Vorhersage positiv ist. Ebenso verhält es sich mit richtig negativ: Der wahre Wert ist negativ und die Vorhersage ebenfalls.

Einen Fehler erster Art erhält man hingegen, wenn der wahre Wert zwar falsch ist, die Vorhersage aber positiv (FP). Die letzte Kombination, der wahre Wert ist richtig und die Vorhersage negativ, beschreibt dann einen Fehler zweiter Art (FN).

Tabelle 1: Konfusionsmatrix.

	Wahr Positiv	Wahr Negativ
Vorhersage Positiv	Richtig Positiv (TP)	Falsch Positiv (FP)
Vorhersage Negativ	Falsch Negativ (FN)	Richtig Negativ (TN)

Aus dieser Matrix ergeben sich verschiedene Qualitätsmetriken, beispielsweise eine True-Positive-Rate, False-Positive-Rate oder die Genauigkeit des Modells. Die Fairness-Maße, die auf dieser Matrix basieren, nutzen diese Qualitätsmetriken und werden mit einer Slicing Analysis errechnet (Gardner et al., 2019; Verma & Rubin, 2018). Dafür wird die Qualitätsmetrik für beide demographische Gruppen separat errechnet und anschließend miteinander verglichen. Wenn der Unterschied zwischen den zwei Ergebnissen einen Schwellenwert überschreitet, dann wird das geprüfte Machine Learning (ML)-Modell als *biased* betrachtet. Verma und Rubin (2018) beschreiben 20 verschiedene Metriken, die auf der Konfusions-Matrix basieren und die Slicing Analysis nutzen. Der Schwellenwert unterscheidet sich je nach angewandter Metrik und Studie und befindet sich zwischen 0.01 und 0.05 (Chouldechova, 2017). Anhand einer Fallstudie konnten Verma und Rubin (2018) zeigen, dass das Modell auf Basis von manchen Maßen als *biased* zu klassifizieren ist, während dem auf Basis von anderen Maßen nicht so ist. Damit zeigt sich die erste Schwierigkeit bei der Bewertung von Systemen: Je nach Auswahl des Fairness-Maßes ergeben sich verschiedene Bewertungen. Eine weitere Schwierigkeit liegt darin, dass für das Testen von Bias nicht nur die vom Modell getroffenen Vorhersagen vorliegen müssen, sondern auch die wahren Werte (Verma und Rubin, 2018). Eine typische Anwendung von ML-Modellen in Universitäten ist die Vorhersage von Durchfallquoten, indem das System voraussagt, ob eine studierende Person

diesen Kurs besteht oder nicht. Durch Bewertung des Systems muss dafür nicht nur das Ergebnis des Prognosemodells vorliegen, sondern auch die Information, ob die studierende Person den Kurs bestanden hat oder nicht. Erst damit kann man herausfinden, ob das Modell falsch lag oder nicht. Das ist zwar meist für Trainingsdaten der Fall, darüber hinaus gibt es aber viele Anwendungsfälle, in denen die wahren Werte nicht so einfach herauszufinden sind.

Es gibt folglich nicht das eine richtige Maß, um Bias in LA-Systemen oder Modellen zu bewerten. Die Auswahl der Metrik muss im Einzelfall und in Abhängigkeit davon entschieden werden, was bewertet werden soll.

2.3 Die Gruppe

Eine weitere Schwierigkeit bei der Auswertung von ML-Modellen oder Lernsystemen liegt in der Auswahl der demographischen Gruppe, über die die Messung etwas aussagen soll. Baker und Hawn (2021) kritisieren hier die Praxis, dass bei der Untersuchung von Diskriminierung oftmals nur nach Geschlecht oder Ethnie (im angloamerikanischen Sprachraum *race*) durchgeführt wird. Es gibt jedoch weitaus mehr Einflüsse der gruppenbezogenen Merkmale, die in Bildungskontexten betrachtet werden sollten. Dazu gehören beispielsweise der sozio-ökonomische Status der Familie (Litman et al., 2021; Yu et al., 2021), der Bildungshintergrund der Eltern (Kai et al., 2017; Rzepka et al., 2022; Yu et al., 2021), die Erstsprache der Lernenden (Loukina et al., 2019; Rzepka et al., 2022) oder auch das Vorliegen einer Behinderung (Loukina & Buzick, 2017; Riazy et al., 2020). Weiterhin kann alleine die Lage der Schule (beispielsweise hinsichtlich Urbanität oder sozialen Lagen) eine Rolle spielen (Ocumpaugh et al., 2014), ebenso wie die Kompetenz der Schüler:innen (Barla et al., 2010). So kann es sein, dass bei der Evaluation eines Systems zwar das Diskriminierungsrisiko in Bezug auf eine Gruppe bestätigt wird, für eine andere Gruppe aber keine diskriminierenden Tendenzen festzustellen sind. Wird der Bias in Bezug auf mehrere Gruppen geprüft, werden diese Gruppen meist separiert betrachtet. Dabei wird meistens nicht auf intersektionale Zusammenhänge geprüft, obgleich Menschen, die in mehrere weniger privilegierte Gruppen fallen, häufig größerer Diskriminierung ausgesetzt sind (Cabrera et al., 2019; Guo & Caliskan, 2021).

2.4 Die Ursache

Sofern eine ›Schwachstelle‹ im LA-System in Bezug auf ein Diskriminierungsrisiko gefunden wurde, ist zu prüfen, an welcher Stelle im Prozess der Bias entstanden ist. Wenngleich oft beschrieben wird, dass das ML-Modell an sich diskriminierend ist, so gibt es noch viele verschiedene Stellen im Erstellungsprozess des Systems, die zur Diskriminierung einer Gruppe führen können (Mehrabi et al., 2021; Mitchell et al.,

2021; Schwartz et al., 2022; Suresh & Guttag, 2021; van Giffen et al., 2022). Suresh und Guttag (2021) haben daher in ihrer Arbeit *7 Sources of Harm* zusammengefasst:

1. Ein *Historical Bias* tritt auf, wenn das ML-Modell mit korrekten, ausgewogenen und ›sauberen‹ Daten trainiert wird, die Daten aber strukturelle Diskriminierung aus der Realität widerspiegeln (Suresh & Guttag, 2021). Würde beispielsweise ein Vorhersagemodell eingesetzt werden, um die Leistung von Studierenden zu prognostizieren, dann könnte hier ein Historical Bias vorliegen. Wenn im analogen Studium Studierende mit internationaler Geschichte von Lehrkräften schlechter bewertet werden als Studierende ohne, dann nutzt das Modell Daten, in denen die strukturelle Diskriminierung bereits enthalten ist. Es kann dann zu Diskriminierung von Studierenden mit Migrationshintergrund kommen.
2. *Representation Bias* hingegen bedeutet, dass bei der Datenerhebung eine demographische Gruppe nicht oder zu wenig berücksichtigt wurde und so in den Daten nicht repräsentiert ist (Suresh & Guttag, 2021). Das kann beispielsweise der Fall sein, wenn man Daten in einer Schule sammelt, in der wenige Kinder mit Migrationshintergrund sind. Beim Einsatz des Modells in dieser Schule würde es für diese Minderheit schlechter funktionieren, weil zu wenige Daten vorhanden sind, um das Modell entsprechend zu trainieren.
3. *Measurement Bias* kann durch eine undurchdachte Wahl und Erfassung von Variablen, die im Modell genutzt werden, entstehen (Suresh & Guttag, 2021). Dies geschieht unter anderem, wenn eine Variable die auszusagende Wirklichkeit stark vereinfacht. Ferner kann *Measurement Bias* entstehen, wenn die Erfassung oder die Genauigkeit der Variable zwischen den verschiedenen Gruppen variiert. Das ist beispielsweise der Fall, wenn eine demographische Gruppe im Schulkontext häufiger falsch bewertet wird als der Durchschnitt. Werden diese Bewertungen im Modell als Inputdaten der Leistung der Schüler:innen genutzt, so sind die Ergebnisse für eine demographische Gruppe weniger akkurat.
4. Ein *Aggregation Bias* entsteht, wenn ein Modell für verschiedene Gruppen genutzt wird, obwohl eine separate Lösung pro Gruppe die Wirklichkeit besser beschreiben würde (Suresh & Guttag, 2021). Dies kann beispielsweise bei der Implementierung eines ML-Modells in verschiedenen Sprachkursen der Fall sein: Je nach Zielgruppe des Sprachkurses unterscheiden sich die Lernenden stark. Gibt es beispielsweise sowohl Sprachkurse für Geflüchtete als auch für die professionelle Weiterbildung im beruflichen Kontext, würden sich bestimmte Prognosen im Mittel zwischen beiden Kursen unterscheiden. Damit wäre aber ein Modell für keine der Gruppen optimal.
5. Ein *Learning Bias* beschreibt die Diskriminierung, die durch das ML-Modell selbst entstehen kann (Mehrabi et al., 2021; Suresh & Guttag, 2021). Bei der Entwicklung des Modells kann typischerweise definiert werden, welche Variable das Modell beim Training optimieren soll. Hooker et al. (2020) zeigten, wie sich

die Entscheidung für ein kompaktes Modell (durch pruning) negativ auf den Bias für unterrepräsentierte Gruppen auswirken kann. Je nach Ausbalancierung der Daten kann es auch hilfreich sein, im Modell auf Precision oder Recall zu optimieren anstelle der Accuracy. Damit würde man der unterrepräsentierten Gruppe eine größere Wichtigkeit im Optimierungsprozess des Modells zuschreiben.

6. *Evaluation Bias* tritt bei der Evaluation des Modells auf (Suresh & Guttag, 2021). Wird hier ein Testdatenset verwendet, das Minoritäten nicht gut repräsentiert, dann wird dem Modell eine Güte zugeschrieben, die nur für die Mehrheit gilt. Wird beispielsweise ein Modell in einer höheren Klassenstufe evaluiert, so kann die attestierte Güte des Modells bei der Anwendung in einer niedrigeren Klassenstufe nicht mehr zutreffen.
7. Und zuletzt tritt ein *Deployment Bias* auf, wenn das Modell in einem Kontext eingesetzt wird, für den es nicht erstellt wurde (Suresh & Guttag, 2021). Ein Beispiel hierzu wäre ein Modell, dass Schüler:innen in Lerntypen gruppieren soll, um die Lehre entsprechend anzupassen. Würden die Ergebnisse der Gruppierung dann aber zur Benotung genutzt, dann würde das Modell nicht zweckmäßig gebraucht und kann zu Diskriminierung führen.

Bei der Vermessung des Lernens und der anschließenden Evaluation von Learning Analytics und adaptiven Lernsystemen gibt es demnach viele Handlungsempfehlungen im gesamten Prozess, die zu Diskriminierung führen können. Insbesondere der hier beschriebenen Vielschichtigkeit von Bias in Lernsystemen wird in der Forschung bislang noch nicht Rechnung getragen, da es für die meisten Studien schwierig ist, sich auf mehrere Maße, mehrere demokratische Gruppen und mehrere Ursachen gleichzeitig zu fokussieren. Während die Forschung zu technischen Möglichkeiten in LA und adaptivem Lernen immer weiter fortschreitet, wird die Fairness dieser Systeme oftmals nur als kleine Erweiterung gesehen – und erreicht damit nicht die Tiefe, die für eine echte Evaluation auf das Diskriminierungsrisiko nötig wäre.

3. Sieben Handlungsempfehlungen zur Sicherstellung von Fairness

Betrachtet man die Vermessung des Lernens und dessen Diskriminierungsrisiko vor dem Hintergrund dieser Ausführungen erneut, so ist vor allem eines deutlich geworden: Die Untersuchung auf Diskriminierung hin ist keine kleine Erweiterung der Entwicklung von Lernsystemen, die zum Ende einmal stattfindet. Vielmehr muss es sich um einen integralen Bestandteil bei der Konzeption dieser Systeme handeln. Das Risiko diskriminierender Systeme besteht in der Verstärkung bestehender Diskriminierung und Ungleichbehandlung von Lernenden. Damit

einher gehen ethische Bedenken, solche Systeme überhaupt einzusetzen, was auch zu einer geringeren Akzeptanz führen kann, sodass mögliche Vorteile solcher Systeme nicht genutzt werden können.

Abgeleitet aus den oben beschriebenen Problemfeldern und bestehender Literatur werden im Folgenden sieben Handlungsempfehlungen vorgestellt und zusammengefasst, die bei der Erstellung und Überprüfung von Lernsystemen beachtet werden sollten:

3.1 Kritische Betrachtung der verwendeten Daten

Bereits vor Beginn des Implementierungsprozesses muss die Integrität der verwendeten Daten kritisch betrachtet werden: Unter welchen Umständen werden die Daten erfasst? Werden in den Daten alle relevanten demographischen Gruppen repräsentiert? Spiegeln die Variablen das wider, für das sie im ML-Modell genutzt werden, oder werden Zusammenhänge in Variablen stark vereinfacht? Es empfiehlt sich daher, schon vor der Auswahl der Trainingsdaten und der Konzeption des Modells und des Lernsystems diese Fragen zu beantworten.

3.2 Welche Metriken sind sinnvoll zur Evaluation meines Lernsystems?

Je nach Aufbau und Zweck des Lernsystems ist darüber hinaus zu überprüfen, mit welchen Metriken man das System am besten evaluieren kann (Verma & Rubin, 2018). Hier kann es sinnvoll sein, verschiedene Maße zu vergleichen und zu überlegen, welche praktischen Auswirkungen als unfair einzustufende Werte von dieser Metrik tatsächlichen hätten. Hierbei ist auch zu prüfen, welche Auswirkungen ein Fehler erster oder zweiter Art auf die Lernerfahrung haben und wie hoch das Diskriminierungsrisiko ist. Wird in einem Lernsystem lediglich implementiert, dass Nutzende mit einer Risikoeinstufung ein ausführlicheres Feedback erhalten, dann ist der Fehler 1. Art schlimmer als der Fehler 2. Art: Erhält ein User, der eigentlich gut genug ist, ein ausführlicheres Feedback (Fehler 2. Art), dann hat das keine großen Auswirkungen auf den Lernerfolg. Wenn jedoch ein User schlechter ist als vom System eingestuft (Fehler erster Art), dann würde er keine ausführlichen Erklärungen erhalten, die ihm möglicherweise helfen würden. Somit würden in diesem Fall Metriken ausgewählt werden, die besonders sensibel auf die False-Positiv-Rate reagieren (beispielsweise die Metrik predictive equality, Verma & Rubin, 2018).

3.3 Einbezug verschiedener demographischen Gruppen

Zur Evaluation des Bias eines Systems ist darüber hinaus zu untersuchen, welche potenzielle Diskriminierung im Kontext des Systems möglich ist (Baker & Hawn,

2021). Meist gibt es hier mehrere demographische Gruppen bzw. Variablen, die Einflussfaktoren sein können. Dabei ist zu berücksichtigen, auf welche demographischen Gruppen das System angewendet wird. Weiterhin kann die Recherche zum Stand der Forschung der jeweiligen Disziplin helfen, um herauszufinden auf Basis welcher Faktoren Lernende diskriminiert werden können. So korreliert beispielsweise im Kontext von Bildung nicht nur das Geschlecht und der Migrationshintergrund mit dem Lernerfolg, sondern auch der sozio-ökonomische Status des Haushalts. Allerdings ist hier ein Trade-Off zwischen Datenschutz und Einbezug demographischer Gruppen zu betrachten. Um demographische Gruppen einzubeziehen und Fairness im Hinblick auf diese Gruppen zu bewerten, muss die Gruppe zunächst bekannt sein. Das bedeutet, dass Daten zu Migrationshintergrund, Geschlecht und/oder sozio-ökonomischem Status erfasst, gespeichert und verarbeitet werden müssen. Die Verarbeitung dieser sensiblen Daten muss dabei zunächst datenschutzrechtlich bewertet werden und die Nutzenden müssen die Daten bereitstellen und der Verarbeitung zustimmen.

3.4 Intersektionale Analyse

Die Analyse verschiedener demographischer Gruppen sollte dabei nicht isoliert stattfinden (Cabrera et al., 2019; Guo & Caliskan, 2021). So kann die Kombination von zwei Diskriminierungskategorien zu weiterer Diskriminierung führen. Zudem kann es vorkommen, dass bei der Überprüfung von einer demographischen Gruppe der Schwellenwert, um von Diskriminierung zu sprechen, nicht erreicht wird, in der Kombination zweier Diskriminierungskategorien aber schon. Daher ist es ratsam, sämtliche demographische Gruppen auch kombiniert auf Diskriminierung zu prüfen, also eine intersektionale Analyse durchzuführen.

3.5 Die Evaluation des Systems über die gesamte Entwicklung des ML-Modells

Grundsätzlich sollte die Evaluation des Systems über die gesamte Entwicklung stattfinden, da wie im obigen Abschnitt beschrieben, verschiedenartige Probleme auftreten können, die zu Diskriminierung führen (Suresh & Guttag, 2021). Die gesamte Entwicklung eines ML-Modells besteht aus fünf Phasen (Yang et al., 2021): (1) Datenmanagement, (2) Modell trainieren, (3) Modell testen, (4) Modell bereitstellen und (5) die Nutzung und Überwachung. In der ersten Phase werden die Daten gesammelt, bereinigt und in Test- und Trainingsdaten unterteilt. Anhand der Trainingsdaten werden verschiedene Modelle ausgewählt, konfiguriert und trainiert. Aufbauend auf das Training der Modelle werden diese getestet. In Phase vier wird das ML-Modell in die Live-Anwendung integriert. Bei erfolgreichem Deployment folgt die letzte Phase, die Nutzung und Beobachtung. Dabei werden alle Ergebnisse

bezüglich der Modelle aufgenommen und für spätere oder anschließende Projekte gespeichert. Die Evaluation muss daher schon mit Beginn der Systementwicklung starten, sodass beispielsweise beim Sammeln der Trainingsdaten bereits ein *Representation Bias* geprüft und ausgeschlossen werden kann. So kann in jedem Entwicklungsschritt die Evaluation der Fairness mitgedacht werden. Weiterhin ist zu beachten, dass auch nach der Implementierung und Bereitstellung des Systems der ML-Lebenszyklus noch nicht beendet ist. Ein ML-Modell muss auch nach der Bereitstellung kontinuierlich überprüft und angepasst werden. Demnach ist auch hier eine Re-Evaluierung der Fairness sinnvoll.

3.6 Einbeziehung der Stakeholder

Für die Bereitstellung eines LA-Systems muss allen Stakeholdern klar sein, für welchen Zweck das System eingesetzt werden soll – und für welche Zwecke nicht. Stakeholder sind dabei alle, die von diesem System direkt oder indirekt betroffen sind. Dazu gehören insbesondere Lernende und Lehrende, aber auch ggf. Eltern, Systemeigner, oder die Institution an sich. Stakeholder sollten darüber aufgeklärt werden, zu welchem Zweck dieses System eingesetzt wird, welche Variablen in die Berechnung einfließen, welches die vorhersagende Variable ist und welche Interventionen bei welchen Werten ausgelöst werden. Die genaue Definition des Einsatzzweckes verhindert den *Deployment Bias* und eine nicht gewollte oder nicht intendierte Nutzung des Lernsystems (Olteanu et al., 2019).

3.7 Transparenz der Systeme schaffen

Je größer der Einfluss von LA und adaptive Learning auf das Lernsystem ist, desto transparenter sollte das System für alle Stakeholder sein. Die Erklärung der zugrundeliegenden Systemkomponenten führt dabei nicht nur zur besseren Akzeptanz, sondern hilft Lehrkräften auch bei der Interpretation der Ergebnisse. Die Befähigung von Lehrkräften und Tutor:innen liefert zudem eine wichtige zusätzliche Kontrollinstanz. Lehrkräfte, die die Grundzüge des Systems nachvollziehen können, stellen falsche oder diskriminierende Rückmeldung schneller in Frage und können das System besser bewerten. Hierfür eignen sich Handlungsempfehlungen von Explainable AI (Lundberg et al., 2020). Unter Explainable AI versteht man ML-Modelle, deren Ergebnisse für den Menschen nachvollziehbar sind (Gunning et al., 2019). Das bedeutet, dass transparent ist, wie das Modell zu dem einen oder anderen Ergebnis kommt und warum.

4. Zusammenfassung und Ausblick

Dieser Beitrag hat zusammengefasst, welche Problemfelder bei der Evaluation von Fairness noch bestehen: die Fairness zu messen, die korrekte Ursache festzustellen und allen demographischen Gruppen gerecht zu werden. Davon ausgehend wurden sieben Leitpunkte in Form von Handlungsempfehlungen entwickelt, die bei der Evaluation eines LA-Systems oder adaptiven Lernsystems zu beachten sind.

Allein die Bewusstmachung der Problemfelder und der Lücken in der Auditierung können helfen, diese Systeme in Zukunft besser auszuwerten. Eine Evaluation von Fairness wendet nicht nur eine spezifische Methode an, sondern sie benötigt ein Methodenspektrum: Je mehr Perspektiven bei der Evaluation eingenommen werden, desto eher wird die Auswertung der Fairness allen Stakeholdern gerecht. Solange die Evaluation von Lernsystemen auf Fairness lediglich als kleine Zusatzaufgabe angesehen wird, werden die dort ermittelten Ergebnisse immer nur eine limitierte Einsicht liefern.

In Zukunft sollte es daher zum Standard werden, ML-Systeme, die implementiert werden, auf deren Fairness zu überprüfen. In Projekten müssen für diese Evaluationen Arbeitsstunden und -pakete entlang der gesamten Projektlaufzeit eingerechnet werden. Gleichzeitig darf die Diskussion um Fairness und Diskriminierung von KI-Systemen nicht dazu führen, dass Ängste und Vorbehalte von Stakeholdern steigen: Denn intelligente Lernplattformen können auch Vorteile für Schüler:innen bringen. Eine umfängliche Evaluation des Diskriminierungsrisikos kann dazu beitragen, dem Ziel eines fairen Lernsystems näher zu kommen. Es kann außerdem die Akzeptanz von Stakeholdern – die oft bereits Vorbehalte haben – erhöhen und so schrittweise einen Beitrag zur Nutzung der technischen Möglichkeiten im Bildungsbereich leisten.

Literatur

Anderson, H., Boodhwani, A., & Baker, R. S. (2019). Assessing the Fairness of Graduation Predictions. *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*. http://radix.www.upenn.edu/learninganalytics/ryanbaker/edm2019_paper56.pdf (zuletzt abgerufen 23.06.2023)

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue. In S. Dawson (Hg.), *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (S. 267). ACM. https://doi.org/10.1145/2330601.2330666 (zuletzt abgerufen 23.06.2023)

Baker, R. S., & Hawn, A. (2021). Algorithmic Bias in Education. *International Journal of Artificial Intelligence in Education*, 1–41.

Barla, M., Bieliková, M., Ezzeddinne, A. B., Kramár, T., Šimko, M., & Vozár, O. (2010). On the impact of adaptive test question selection for learning efficiency. *Computers & Education*, *55*(2), 846–857.

Bodily, R., & Verbert, K. (2017). Review of Research on Student-Facing Learning Analytics Dashboards and Educational Recommender Systems. *IEEE Transactions on Learning Technologies*, *10*(4), 405–418.

Cabrera, A. A., Epperson, W., Hohman, F., Kahng, M., Morgenstern, J., & Chau, D. H. (2019). FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)* (S. 46–56). IEEE. doi.org/10.1109/VAST47406.2019.8986948

Chouldechova, A. (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, *5*(2), 153–163.

Duden (2022). *Fairness, die*. https://www.duden.de/rechtschreibung/Fairness (zuletzt abgerufen 23.06.2023)

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In S. Goldwasser (Hg.), *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on – ITCS '12* (S. 214–226). ACM Press.

Ebner, M. (2019). Learning Analytics. Eine Einführung. *Bildung und Beruf*, *2*(2), 46–49.

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

Griff, E. R., & Matter, S. F. (2013). Evaluation of an adaptive online learning system. *British Journal of Educational Technology*, *44*(1), 170–176.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.Z. (2019). Xai-Explainable artificial intelligence. *Science Robotics*, *4*(37).

Guo, W., & Caliskan, A. (2021). Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. In M. Fourcade (Hg.), *ACM Digital Library, Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (S. 122–133). Association for Computing Machinery. doi.org/10.1145/3461702.3462536

Hooker, S., Moorosi, N., Clark, G., Bengio, S., & Denton, E. (2020). *Characterising Bias in Compressed Models*. https://arxiv.org/pdf/2010.03058 (zuletzt abgerufen 23.06.2023)

Hooshyar, D., Yousefi, M., & Lim, H. (2018). A Procedural Content Generation-Based Framework for Educational Games: Toward a Tailored Data-Driven Game for Developing Early English Reading Skills. *Journal of Educational Computing Research*, *56*(2), 293–310.

Hu, Q., & Rangwala, H. (2020). Towards Fair Educational Data Mining: A Case Study on Detecting At-risk Students. *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)*, 431–437.

Jeong, H., Wu, M., Dasgupta, N., Medard, M., & Calmon, F. (2022). *Who Gets the Benefit of the Doubt? Racial Bias in Machine Learning Algorithms Applied to Secondary School Math Education*. https://fated2022.github.io/assets/pdf/fated-2022_paper_jeong_racial_bias_ml_algs.pdf (zuletzt abgerufen 23.06.2023)

Kai, S., Andres, J. M. L., Paquette, L., Baker, R. S., Molnar, K., Watkins, H., & Moore, M. (2017). Predicting Student Retention from Behavior in an Online Orientation Course. *International Educational Data Mining Society*. https://eric.ed.gov/?id=ed596601 (zuletzt abgerufen 23.06.2023)

Kizilcec, R. F., & Lee, H. (2020). Algorithmic Fairness in Education. *The Ethics of Artificial Intelligence in Education*, *8*(11).

Klare, B. F., Burge, M. J., Klontz, J. C., Vorder Bruegge, R. W., & Jain, A. K. (2012). Face Recognition Performance: Role of Demographic Information. *IEEE Transactions on Information Forensics and Security*, *7*(6), 1789–1801.

Lipton, Z., McAuley, J., & Chouldechova, A. (2018). Does mitigating ML\textquotesingle s impact disparity require treatment disparity? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Hg.), *Advances in Neural Information Processing Systems* (Vol. 31). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/file/8e0384779e58ce2af40eb365b318cc32-Paper.pdf (zuletzt abgerufen 23.06.2023)

Litman, D., Zhang, H., Correnti, R., Matsumura, L. C., & Wang, E. (2021). A Fairness Evaluation of Automated Methods for Scoring Text Evidence Usage in Writing. *Lecture Notes in Computer Science*, Volume 12748, 255–267,

Loukina, A., & Buzick, H. (2017). Use of Automated Scoring in Spoken Language Assessments for Test Takers With Speech Impairments. *ETS Research Report Series*, *2017*(1), 1–10.

Loukina, A., Madnani, N., & Zechner, K. (2019). The many dimensions of algorithmic fairness in educational applications. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics. doi.org/10.18653/v1/w19-4401

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.I. (2020). From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*, *2*(1), 56–67.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6).

Meier, C. (2019). KI-basierte, adaptive Lernumgebungen. In K. Wilbers (Hg), *Handbuch E-Learning* (S. 1–21). Deutscher Wirtschaftsdienst / Luchterhand / Wolters Kluwer.

Mirata, V., Hirt, F., Bergamin, P., & van der Westhuizen, C. (2020). Challenges and contexts in establishing adaptive learning in higher education: findings from

a Delphi study. *International Journal of Educational Technology in Higher Education*, *17*(1).

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, *8*(1), 141–163.

Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. (2014). Population validity for educational data mining models: A case study in affect detection. *British Journal of Educational Technology*, *45*(3), 487–501.

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data*, *2*(13).

Paramythis, A., & Loidl-Reisinger, S. (2003). Adaptive learning environments and e-learning standards. In R. Williams (Hg.), *2nd European Conference on e-Learning: Glasgow Caledonian University, Glasgow, 6–7 November 2003*. Academics Conferences International.

Penn Center for Learning Analytics. (2022). *Empircal Evidence for Algorithmic Bias in Education: The Wiki.* https://www.pcla.wiki/index.php/Algorithmic_Bias_in_Education (zuletzt abgerufen 23.06.2023)

Riazy, S., & Simbeck, K. (2019). *Predictive Algorithms in Learning Analytics and their Fairness*. 1617–5468. Advance online publication. doi.org/10.18420/delfi2019_305

Riazy, S., Simbeck, K., & Schreck, V. (2020). Fairness in Learning Analytics: Student At-risk Prediction in Virtual Learning Environments. In *Proceedings of the 12th International Conference on Computer Supported Education – Volume 1: CSEDU* (S. 15–25). SCITEPRESS.

Rzepka, N., Simbeck, K., Müller, H. G., & Pinkwart, N. (2022). Fairness of In-session Dropout Prediction. In *Proceedings of the 14th International Conference on Computer Supported Education* (Vol. 2, S. 316–326). SCITEPRESS. doi.org/10.5220/0010962100003182

Scherr, A. (2020). Diskriminierung und Diskriminierungskritik: eine problemsoziologische Analyse. *Soziale Probleme*, *31*(1–2), 83–102.

Scherr, A., El-Mafaalani, A., & Yüksel, G. (2017). *Handbuch Diskriminierung*. Springer Fachmedien Wiesbaden. doi.org/10.1007/978-3-658-10976-9

Schiepe-Tiska, A., Köller, O., Sälzer, C., Klieme, E., & Reiss, K. (2016). *PISA 2015*. Waxmann Verlag. https://directory.doabooks.org/handle/20.500.12854/56295 https://doi.org/56295 (zuletzt abgerufen 23.06.2023)

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence.* doi.org/10.6028/NIST.SP.1270

Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, *27*(12), 2176–2182.

Siemens, G. (2013). Learning Analytics. *American Behavioral Scientist*, *57*(10), 1380–1400.

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

Sweeney, L. (2013). Discrimination in online ad delivery. *Commun. ACM*, *56*(5), 44–54.

Title VII of the Civil Rights Act of 1964. https://www.eeoc.gov/statutes/title-vii-civil-rights-act-1964 (zuletzt abgerufen 23.06.2023)

Valtin, R., Badel, I., Löffler, I., Meyer-Schepers, U., & Voss, A. (2003). *Orthographische Kompetenzen von Schülerinnen und Schülern der vierten Klasse*. Waxmann. doi.org/10.25656/01:14854

van Giffen, B., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research*, *144*, 93–106.

van Oostendorp, H., van der Spek, E. D., & Linssen, J. (2014). Adapting the Complexity Level of a Serious Game to the Proficiency of Players. *EAI Endorsed Transactions on Game-Based Learning*, *1*(2), e5.

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

Yang, C., Wang, W., Zhang, Y., Zhang, Z., Shen, L., Li, Y., & See, J. (2021). Mlife: A lite framework for machine learning lifecycle initialization. *Machine Learning*, *110*(11–12), 2993–3013.

Yu, R., Lee, H., & Kizilcec, R. F. (2021). Should College Dropout Prediction Models Include Protected Attributes? In C. Meinel (Hg.), *ACM Digital Library, Proceedings of the Eighth ACM Conference on Learning Scale* (S. 91–100). Association for Computing Machinery. doi.org/10.1145/3430895.3460139

Yu, R., Li, Q., Fischer, C., Doroudi, S., & Di Xu (2020). Towards Accurate and Fair Prediction of College Success: Evaluating Different Sources of Student Data. *International Educational Data Mining Society*. https://eric.ed.gov/?id=ed608066 (zuletzt abgerufen 23.06.2023)