Marie Decker, Carmen Leicht-Scholten

„Intersektionale Algorithmen?" Herausforderungen und Chancen der Operationalisierung von Intersektionalität

Marie Decker (Foto: Bettina Steinacker).

1 Einleitung

Als Kimberlé Crenshaw im Jahr 1989 den Begriff *Intersektionalität* prägte (Crenshaw 1989), kritisierte sie, wie die separate Betrachtung von Gender und Race im (US-amerikanischen) Antidiskriminierungsgesetz dazu führte, dass Schwarzen[1] Frauen das Repräsentationsrecht sowohl für Frauen als auch für Schwarze Personen aberkannt wurde. Diskriminierung würde nur für die privilegiertesten Individuen evaluiert und mehrfach diskriminierte Personen weiter marginalisiert. Und auch nach 35 Jahren breiter Diskurse von Intersektionalität (Collins 2015) finden wir noch heute viele der von Crenshaw diskutierten Aspekte im Kontext von Algorithmischer Fairness explizit wieder.

Algorithmische Fairness im Kontext dieses Artikels bezieht sich auf die mathematische Fairness-Evaluation von algorithmischen Entscheidungssystemen (kurz: ADM), die in verschiedensten Bereichen der Gesellschaft Anwendung finden: zur Entscheidung über die Kreditwürdigkeit von Personen (Gikay 2020), deren Strafmaß (Berk et al. 2018; Wang 2018) oder zur Entscheidung über Unterstützung bei Jobentscheidungen oder -fördermaßnahmen (Allhutter et al. 2020; Raghavan et al. 2020).

Juristische und algorithmische Entscheidungsfindungen weisen einige Parallelen auf: Auch wenn im traditionellen Rechtssystem menschliche Jurist*innen die Auslegung von Gesetzen vornehmen, so folgen beide Arten der Entscheidungsfindung klaren Regeln, streben nach Vergleichbarkeit und zielen auf eine Form von Neutralität ab (Sundar/Kim 2019). Die „aura of truth, objectivity, and accuracy" (Boyd/Crawford 2012: 663) algorithmischer Systeme ist jedoch irreführend: Die Relevanz von Artefakten liegt einzig bei den genutzten Daten und implementierten Algorithmen selbst. In der Folge wurde schon früh die Reproduktion gesellschaftlicher Stereotype und Herrschaftsverhältnisse durch den Einsatz von ADM bekannt und diskutiert (Angwin et al. 2016; Barocas/Selbst 2016).

Als (technische) Reaktion darauf entstanden zahlreiche mathematische Fairness-Metriken, die durch einen meist distributiven Ansatz (Kasirzadeh 2022) eine Form der Gleichbehandlung von marginalisierten Gruppen sicherstellen sollen (Barocas/Hardt/Narayanan 2023; Corbett-Davies/Goel 2018; Mehrabi et al. 2019; Pessach/Shmueli 2022). Dabei werden rechtlich geschützte Kategorien wie Race oder Gender, seltener Alter, Familienstatus oder Ethnizität (Simson/Fabris/Kern 2024) hinsichtlich paritätischer Ergebnisse evaluiert. Diese eindimensionale Betrachtung, „to treat race and gender as mutually exclusive categories of experience and analysis", kritisierte bereits Crenshaw (1989: 139). Einige technische Pionierarbeiten erweitern die Fairness-Metriken um die Betrachtung von Sub-Kategorien und bezeichnen dies als *Intersektionale Algorithmische Fairness.* Dieser Begriff ist jedoch irreführend, da Intersektionalität neben einer multidimensionalen Betrachtung von Diskriminierungserfahrungen die vielfältig miteinander verwobenen Prozesse von Privilegierung und Diskriminierung beschreibt, die strukturelle Diskriminierung aufrechterhalten (Collins 2019). Dies zeigt, dass Intersektionalität als sozialwissenschaftliches Konzept im Kontext der Algorithmischen Fairness bislang nur unzureichend oder missverständlich adressiert wird (Ovalle et al. 2023).

[1] Schwarz wird in diesem Artikel intentional großgeschrieben, um die politische Bedeutung der Kategorie hervorzuheben.

In diesem Artikel erläutern wir zunächst aktuelle Konzeptionen von Intersektionalität im Kontext Algorithmischer Fairness. Anschließend identifizieren wir fünf spezifische Herausforderungen, die die (technische) Operationalisierung von Intersektionalität im sozialwissenschaftlichen Sinne derzeit erschweren. Abschließend diskutieren wir, warum es dennoch wertvoll ist, diesen Ansatz weiterzuverfolgen, und welche Voraussetzungen dafür geschaffen werden müssen.
Wissen und Wissensbildung sind nicht objektiv – insbesondere im Kontext von Intersektionalität: „Intersectionality faces a particular definitional dilemma – it participates in the very power relations that it examines and, as a result, must pay special attention to the conditions that make its knowledge claims comprehensible." (Collins 2015: 3) Wir Autorinnen dieses Papers sind weiße Frauen aus Deutschland, einem WEIRD-Land (Henrich/Heine/Norenzayan 2010). Unsere kontextuelle Situiertheit als Wissenschaftlerinnen wird also trotz größter Sorgfalt und Selbstreflexion in diesen Artikel mit einfließen.

2 Aktuelle Konzeptionen von Intersektionalität im Kontext Algorithmischer Fairness

Aktuell diskutierte Ansätze zur Evaluation sogenannter Intersektionaler Algorithmischer Fairness sind aus technischen Ansätzen zur Evaluation von *unidimensionaler* Algorithmischer Fairness hervorgegangen. Es wird zwischen drei Ansätzen unterschieden (Barocas/Hardt/Narayanan 2023; Chouldechova/Roth 2020): Individuelle Ansätze fordern ähnliche Vorhersagen für ähnliche Individuen (Dwork et al. 2012; Kusner et al. 2017), kausale Ansätze minimieren den Einfluss von Gruppenzugehörigkeiten auf die Vorhersagen (Kilbertus et al. 2017) und statistische Ansätze sollen eine Form von Parität der Vorhersagen zwischen unterschiedlichen marginalisierten Gruppen herstellen (Caton/Haas 2020). Während individuelle und kausale Ansätze praktischen Umsetzungsschwierigkeiten gegenüberstehen, sind statistische Ansätze weit verbreitet (Bellamy et al. 2018). Aus technischer Perspektive bieten statistische Herangehensweisen viele Vorteile: Durch die Untersuchung von Parität der Vorhersagen müssen keine komplizierten Voraussetzungen an die Trainingsdaten gestellt werden, sodass Algorithmische Fairness recht einfach quantifiziert werden kann.
Zusätzlich zu der Tatsache, dass dieses Vorgehen jedoch voraussetzt, dass zulässige und unzulässige Diskriminierung allein durch die entsprechende Gruppenzugehörigkeit definiert werden kann (Heinrichs 2021; Hoffmann 2019; Lee/Floridi/Singh 2021), wird dabei die soziale Konstruiertheit von Kategorien nicht berücksichtigt (Andrus/Villeneuve 2022; Hanna et al. 2020; Krupiy 2020; Leavy/Siapera/O'Sullivan 2021; Zimmermann/Lee-Stronach 2022). Diese Fixierung auf starre Kategorien kodiert nicht nur gesellschaftliche Normen und vernachlässigt subjektive Erfahrungen, sie konzeptualisiert dynamische Konzepte auch als natürliche und objektive Tatsachen (Green 2022; Leavy/Siapera/O'Sullivan 2021), was schlussendlich Stereotype und Stigmatisierung verfestigen oder gar verstärken kann.
Zusätzlich entstehen Problematiken insbesondere im Kontext einer *intersektionalen* Perspektive (Cho/Crenshaw/McCall 2013; Collins 2015; Collins/Bilge 2016; Crenshaw 1989, 1991; Hancock 2007; Nash 2017). Wie Joy Buolamwini und Timnit Gebru in der Studie „Gender Shades" (Buolamwini/Gebru 2018) anhand von Gesichtserkennungssoftwares verdeutlichten, können Algorithmen für einzelne Kategorien zwar als fair evaluiert werden, für Sub-Kategorien jedoch stark verzerrte Ergebnisse liefern (sog. *Fairness Gerrymandering*). In der Folge sind Arbeiten entstanden, die Algorithmische Fairness als Form der Parität hinsichtlich mehrerer Sub-Kategorien simultan evaluieren (Gohar/Cheng 2023). Es werden also nicht mehr nur z. B. Frauen und Männer oder Schwarze und weiße, sondern nun auch Schwarze und weiße Männer sowie Schwarze und weiße Frauen (als Sub-Kategorien von Gender x Race) betrachtet (Foulds et al. 2018b; Ghosh/Genuit/Reagan 2021; Hebert-Johnson et al. 2018; Islam et al. 2023; Kearns et al. 2018; Kim/Reingold/Rothblum 2018; Yona/Rothblum 2018). Offene Fragestellungen bei diesen Ansätzen beziehen sich damit auf die richtige Anzahl und Auswahl von Sub-Kategorien:
„There are exponentially many ways of carving up a population into subgroups, and we cannot necessarily identify a small number of these a priori *as the only ones we need to be concerned about. At the same time, we cannot insist on any notion of statistical fairness for* every *subgroup of the population: for example, any imperfect classifier could be accused of being unfair to the subgroup of individuals defined ex-post as the set of individuals it misclassified. This simply corresponds to ‚overfitting' a fairness constraint." (Kearns et al. 2018: 2)*
Die Reduktion komplexer Diskriminierungserfahrungen auf Sub-Gruppen (im Folgenden *Sub-Gruppen Fairness* genannt) greift sozialwissenschaftliche Fragestellungen damit nur oberflächlich auf und wird dem Konzept der In-

tersektionalität nicht gerecht (J. L. Davis/Williams/Yang 2021; Hoffmann 2019; Kong 2022; Ovalle et al. 2023). Wie Cho/CrenshawMcCall (2013: 795) betonen, „[w]hat makes an analysis intersectional is not its use of the term 'intersectionality'. […] Rather, what makes an analysis intersectional is its adoption of an intersectional way of thinking about the problem of sameness and difference and its relation to power."

3 Herausforderungen und Potenziale in der Operationalisierung von Intersektionalität

Dieses Kapitel synthetisiert fünf Herausforderungen der technischen Operationalisierung von Intersektionalität: die Konzeptunschärfe, eine fehlende Anerkennung der Vielschichtigkeit von Intersektionalität, eine fehlende kritische Hinterfragung von Machtperspektiven, die Verwendung von Kategorien zur Diskriminierungsevaluation sowie die kontextabhängige Wirkung von Entscheidungen.

3.1 Konzeptunschärfe

Die erste Herausforderung liegt in der inhärenten Unschärfe und Mehrdeutigkeit des Konzeptes von Intersektionalität selbst. Obwohl „[s]cholars and practitioners think they know intersectionality when they see it", wie Collins (2015: 3) feststellt, „they conceptualize intersectionality in dramatically different ways when they use it." Zwar existiert ein weitgehend akzeptierter Orientierungsrahmen, der verschiedene Analyseebenen, soziale Kategorien und Themenfelder miteinander verknüpft[2]. Dennoch besteht Uneinigkeit über eine einheitliche Begriffsbestimmung oder die praktische Anwendung (K. Davis 2008). Intersektionalität kann als Theorie, Konzept, heuristisches Instrument, Lesestrategie für feministische Analyse, Methodik oder Paradigma verstanden werden (Walgenbach 2011, 2012). Somit ist Intersektionalität „not a singular theory, but an approach and a prism with a set of orienting assertions, goals, and tools." (J. L. Davis/Williams/Yang 2021: 3)
Während genau diese Offenheit, Unschärfe und Ambiguität von Intersektionalität für die theoretische und gesellschaftspolitische Diskussion essenziell ist (K. Davis 2008: 67), erschwert sie zugleich eine präzise Operationalisierung im Kontext von Algorithmischer Fairness. Für eine Operationalisierung wäre eine klare Strukturierung nötig; eine Einheitsdefinition wäre jedoch in sich problematisch, denn eine präzise Bestimmung widerspräche der Vielschichtigkeit des Konzeptes: „Presenting a finished definition of intersectionality that can be used to determine whether a given book, article, law, or practice fits within a preconceived intersectional framework misreads […] intersectionality's complexity." (Collins 2015: 3)

3.2 Fehlende Anerkennung der Vielschichtigkeit von Intersektionalität

Während mittlerweile weitgehend anerkannt ist, dass sozialwissenschaftliche Konzepte und Theorien für die Entwicklung von menschenzentrierten algorithmischen Systemen unerlässlich sind, so werden ebenjene Konzepte und Diskurse oft verkürzt. So wird auch das Konzept der Intersektionalität als feststehendes, deterministisches Konzept aufgefasst: „AI research interprets intersectionality as a dimension of ‚solvability' and scale" (Ovalle et al. 2023: 504). In der Folge wird Intersektionalität entweder unkritisch als Sub-Gruppen Fairness übernommen (Kong 2022) oder als komplexes Konzept genannt, aber dann doch als Sub-Gruppen Fairness operationalisiert: „We find that even when researchers center intersectionality literature, there is little engagement with the framework itself, evidenced by a lack of described social context, little discussion of power and relations between structures, questionable citational practices, and a disjointed sense of social justice praxis." (Ovalle et al. 2023: 497) Diese Herangehensweise zeigt sich auch in einer oft sehr vagen Begrifflichkeit zur Einführung von Intersektionalität oder wird durch Re-Zitation weiter verstärkt[3].

3.3 Kritische Hinterfragung von Machtperspektiven

Auch wenn dieses Vorgehen auf eine unzureichende Auseinandersetzung mit sozialwissenschaftlichen Theorien hindeutet, betont Kong (2022: 488): „Most engineers and crowdworkers do not deliberately discriminate against Black women. The problem is rather that they are simply doing their jobs but their actions contribute to reproducing oppression." Hoffmann (2019: 904) führt dies explizit auf systemische Ursachen zurück: „[a]ppealing to the 'blindspots' of particular designers or teams ignores the structuring role of technology, instead reducing a system's shortcomings to the biases of its imperfect human designers." Burrell und Fourcade (2021) prägen in diesem Kontext den Begriff der *coding-elite,* einer neuen ge-

[2] Collins und Bilge (2016) identifizieren etwa fünf zentrale Fokuspunkte: soziale Gerechtigkeit, soziale Ungleichheit, soziale Macht, sozialer Kontext und Komplexität.

[3] Gohar und Cheng (2023) zitieren in ihrem Artikel z. B. andere Arbeiten wie Kearns et al. (2018) oder Hebert-Johnson et al. (2018) in Bezug auf das Konzept der Intersektionalität, obwohl diese präzise und explizit nur von Sub-Gruppen Fairness und nicht von Intersektionalität sprechen.

sellschaftlichen Klasse, die ihre Macht durch technische Kontrolle über digitale Produktionsmittel erworben hat und sich ihrer historischen und sozialen strukturellen Voreingenommenheit nicht zwangsläufig bewusst ist. Nichtsdestotrotz nehmen sie Einfluss auf die Gestaltung von Algorithmen – und damit auf die Gestaltung von Wissen. Zum Beispiel ist die Auswahl der Zielfunktion eines Algorithmus maßgeblich beeinflussend für das Ergebnis und damit – auf weite Sicht – auch für die Verteilung von Ressourcen, und wird damit zu einer höchst politischen Entscheidung[4] (Kasy/Abebe 2021).

Collins und Bilge (2016) beschreiben Intersektionalität als analytisches Werkzeug zur Untersuchung von Machtmechanismen entlang struktureller, disziplinärer, kultureller und zwischenmenschlicher Machtfelder (Collins/Bilge 2016). Eine intersektionale Perspektive auf Algorithmische Fairness erfordert daher, die Perspektive der privilegierten Entwickler*innen und Forscher*innen in den Vordergrund zu rücken, zugrunde liegende Annahmen explizit offenzulegen (Mitchell et al. 2021) und diese mit einer kritischen Reflexivität zu betrachten (Ovalle et al. 2023), beispielsweise mithilfe partizipativer Ansätze (Decker/Wegner/Leicht-Scholten 2025).

3.4 Verwendung von Kategorien zur Diskriminierungsevaluation

Ein zentraler Aspekt von Intersektionalität ist die situative Sichtbarkeit von Kategorien und deren Bedeutung im jeweiligen sozio-strukturellen Kontext. Soziale Identitäten sind keine statischen Attribute, sondern komplexe Konstrukte, die im Wechselspiel mit gesellschaftlichen Strukturen und Institutionen entstehen und nur innerhalb dieses Kontextes verstanden werden können (J. L. Davis/Williams/Yang 2021). Der Fokus auf vordefinierte Kategorien ist aus Sicht von algorithmischen Systemen praktisch, da solche Kategorien leicht operationalisierbar und messbar sind und in datengetriebenen Modellen eine systematische Evaluierung ermöglichen (Alzubi/Nayyar/Kumar 2018).

Wenn sich allerdings die Diskussion zu Intersektionaler Algorithmischer Fairness auf die Auswahl und Kombination von Sub-Kategorien beschränkt, wird Intersektionalität „a matter of splitting a group into finer subgroups along the lines of identity categories" (Kong 2022: 487). In der Folge wird die Auswahl der betrachteten Sub-Kategorien datengetrieben und nicht sozial kontextualisiert getroffen, was bedeutet, dass Kategorien angeordnet und hierarchisiert werden müssen, anstatt die kontextabhängigen Machtverhältnisse zwischen diesen zu beleuchten (Hoffmann 2019). Das führt jedoch dazu, dass vor allem größere, sichtbarere Gruppen in Fairness-Evaluierungen einbezogen werden, während kleinere und historisch marginalisierte Gruppen unterrepräsentiert bleiben (Kong 2022). Weiterhin verstellt die rein datenbasierte Festlegung relevanter Identitätskategorien den Blick auf die politische und soziale Bedeutung der Kategorien und entpolitisiert die Strukturursachen von Diskriminierung (Hoffmann 2019). Denn Intersektionalität ist nicht bloß das zufällige Kombinieren variabler Kategorien, um Benachteiligungen aufzuzeigen; vielmehr verlangt sie eine Analyse der sozialen und historischen Bedingungen, die die Struktur und Kontingenz sozialer Kategorien bestimmen (Hoffmann 2019). Collins (2019) weist darauf hin, dass eine analytisch unreflektierte Gruppenzuweisung sozialen Kategorien ihren ursprünglichen Bedeutungsgehalt entzieht und diese auf „descriptive, non-analytical" Merkmale reduziert. Ovalle et al. (2023) heben hervor, dass nur wenige Arbeiten, die sich mit Intersektionaler Algorithmischer Fairness beschäftigen, sich auch mit den kolonialen Strukturen von geschützten Kategorien beschäftigen oder Gruppenzugehörigkeiten mit sozialen Strukturen verknüpfen. Darüber hinaus blendet der Fokus auf geschützte Kategorien die Erzeugung und Sichtbarmachung von Privilegien aus, indem sie den Fokus primär auf relative Benachteiligungen legt. Hoffmann (2019: 907) betont, dass derartige Ansätze den systematischen Vorteil privilegierter Gruppen ausklammern: „computational solutions to problems of fairness almost exclusively focus on disadvantage. […] The shift is subtle, but consequential: by centering disadvantage, we fail to question the normative conditions that produce—and promote the qualities or interests of—advantaged subjects."

3.5 Kontextabhängige Wirkung von Entscheidungen

Die letzte der hier skizzierten Herausforderungen besteht darin, dass mit den aktuellen technischen Interventionen lediglich an den Outcomes der Algorithmen angesetzt wird[5]. Dies lässt jedoch außer Acht, dass dieselbe Entscheidung unterschiedliche Auswirkungen auf verschiedene Gruppen haben kann (J. L. Davis/Williams/Yang 2021). Ungleichheiten manifestieren sich auf vielfältige Weise und entstehen in Abhängigkeit von legalen, persönlichen und beruflichen Kontexten (J. L. Davis/Williams/Yang 2021).

[4] Beispielhaft führen Ovalle et al. (2023) an, wie Foulds et al. (2018a) die Vorhersage der Rückfallquote nach Straftaten als Maßstab für Fairness verwenden, obwohl in diesem Kontext strukturelle Polizeigewalt diskutiert werden müsste.

[5] Während die aktuellen Ansätze (Intersektionaler) Algorithmischer Fairness auf die faire Verteilung der Vorhersagen des Algorithmus abzielen, so richtet *distributive justice* den Blick auf die Güter, deren Verteilung aus diesen Vorhersagen hervorgeht (Zezulka und Genin 2024).

Die Herausforderung an dieser Stelle besteht darin, Algorithmische Fairness nicht als isoliertes Problem des Algorithmus zu betrachten, sondern die komplexen Systeme zu berücksichtigen, in die Algorithmen eingebettet sind. Hoffmann (2019: 910) schreibt dazu: „critics of distributive conceptions of justice further show that exclusively attending to goods like rights, opportunities, and material resources—while important—are not sufficient for dismantling or upending these hierarchies." Verstärkend wirkt hier, dass die Entwicklung von ADM und deren tatsächliche Einsatzkontexte oft voneinander entkoppelt sind – insbesondere bei sogenannten Multi-Purpose-Systemen, die für eine breite und oft unbekannte Anwendungsbandbreite konzipiert sind. Diese Flexibilität erschwert es, die sozialen Auswirkungen spezifischer Einsatzkontexte im Vorfeld zu analysieren und einzuplanen.

4 Ausblick

In diesem Artikel wurde beleuchtet, dass Intersektionalität in Diskussionen zur Algorithmischen Fairness meist unzureichend berücksichtigt oder verkürzt dargestellt wird. Weiterhin wurden zentrale Herausforderungen aufgezeigt, die bei der Operationalisierung von Intersektionalität auftreten können. Bestehende technische Ansätze, die Ziele Intersektionaler Algorithmischer Fairness unterstützen könnten, sollten in einen breiteren, holistischeren Fairnessdiskurs integriert werden. Ein Wechsel des Schwerpunkts hin zu einer Analyse des Zwecks, Kontexts und gesellschaftlicher Wirkung von Algorithmen kann dabei neue Wege zur Förderung sozialer Gerechtigkeit eröffnen. Die aktuelle Algorithmische-Fairness-Debatte erfordert daher eine stärkere Anlehnung an sozial- und geisteswissenschaftliche Ansätze, insbesondere der feminist science and technology studies, damit Intersektionalität nicht nur als theoretisches Konstrukt, sondern auch als praktisches Werkzeug verstanden werden kann. In zukünftigen Arbeiten wird angestrebt, praxisnahe und inhaltlich fundierte Ansätze für die Implementierung Intersektionaler Algorithmischer Fairness zu entwickeln. Ziel ist es, ein tiefgehendes Verständnis für das Konzept in die Praxis zu übersetzen und Entwickler*innen Werkzeuge an die Hand zu geben, die trotz unvermeidbarer Vereinfachungen den Kern von Intersektionalität bewahren.

Literaturverzeichnis

- Allhutter, Doris; Mager, Astrid; Cech, Florian; Fischer, Fabian & Grill, Gabriel. (2020). *DER AMS-ALGORITHMUS: Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS). Endbericht.* Report No. 2020-02. https://doi.org/10.1553/ITA-pb-2020-02.
- Alzubi, Jafar; Nayyar, Anand & Kumar, Akshi. (2018). Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series,* 1142, 12012. https://doi.org/10.1088/1742-6596/1142/1/012012.
- Andrus, McKane & Villeneuve, Sarah. (2022). Demographic-Reliant Algorithmic Fairness: Characterizing the Risks of Demographic Data Collection in the Pursuit of Fairness. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (S. 1709–1721). ACM. https://doi.org/10.1145/3531146.3533226.
- Angwin, Julia; Larson, Jeff; Kirchner, Lauren & Mattu, Surya. (2016). *Machine Bias.* Zugriff am 25. November 2024 unter www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
- Barocas, Solon; Hardt, Moritz & Narayanan, Arvind. (2023). *Fairness and Machine Learning: Limitations and Opportunities.* MIT Press.
- Barocas, Solon & Selbst, Andrew D. (2016). *Big Data's Disparate Impact.* https://doi.org/10.2139/ssrn.2477899.
- Bellamy, Rachel K. E.; Dey, Kuntal; Hind, Michael; Hoffman, Samuel C.; Houde, Stephanie; Kannan, Kalapriya; Lohia, Pranay; Martino, Jacquelyn; Mehta, Sameep; Mojsilovic, Aleksandra; Nagar, Seema; Ramamurthy, Karthikeyan N.; Richards, John; Saha, Diptikalyan; Sattigeri, Prasanna; Singh, Moninder; Varshney, Kush R. & Zhang, Yunfeng. (2018). *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias.* https://arxiv.org/pdf/1810.01943v1.
- Berk, Richard; Heidari, Hoda; Jabbari, Shahin; Kearns, Michael & Roth, Aaron. (2018). Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research,* 50(1), 3–44. https://doi.org/10.1177/0049124118782533.
- Boyd, Danah & Crawford, Kate. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society,* 15(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878.
- Buolamwini, Joy & Gebru, Timnit. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Sorelle A. Friedler & Christo Wilson (Hrsg.),

Proceedings of Machine Learning Research, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (S. 77–91). PMLR. https://proceedings.mlr.press/v81/buolamwini18a.html.

- Burrell, Jenna & Fourcade, Marion. (2021). The Society of Algorithms. *Annual Review of Sociology,* 47(1), 213–237. https://doi.org/10.1146/annurev-soc-090820-020800.
- Caton, Simon & Haas, Christian. (2020). Fairness in Machine Learning: A Survey. Vorab-Onlinepublikation. https://doi.org/10.48550/arXiv.2010.04053.
- Cho, Sumi; Crenshaw, Kimberlé W. & McCall, Leslie. (2013). Toward a Field of Intersectionality Studies: Theory, Applications, and Praxis. *Signs: Journal of Women in Culture and Society,* 38(4), 785–810. https://doi.org/10.1086/669608.
- Chouldechova, Alexandra & Roth, Aaron. (2020). A Snapshot of the Frontiers of Fairness in Machine Learning. *Commun.* ACM, 63(5), 82–89. https://doi.org/10.1145/3376898.
- Collins, Patricia H. (2015). Intersectionality's Definitional Dilemmas. *Annual Review of Sociology,* 41(1), 1–20. https://doi.org/10.1146/annurev-soc-073014-112142.
- Collins, Patricia H. (2019). *Intersectionality as critical social theory.* Duke University Press. https://doi.org/10.1515/9781478007098.
- Collins, Patricia H. & Bilge, Sirma. (2016). *Intersectionality.* Polity Press. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=4698012.
- Corbett-Davies, Sam & Goel, Sharad. (2018). *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning.* https://arxiv.org/pdf/1808.00023.
- Crenshaw, Kimberlé W. (1989). Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum, 1989*, Artikel 8, 139. https://scholarship.law.columbia.edu/faculty_scholarship/3007.
- Crenshaw, Kimberlé W. (1991). Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review,* 43(6), 1241. https://doi.org/10.2307/1229039.
- Davis, Jenny L.; Williams, Apryl & Yang, Michael W. (2021). Algorithmic reparation. *Big Data & Society,* 8(2), 205395172110448. https://doi.org/10.1177/20539517211044808.
- Davis, Kathy. (2008). Intersectionality as buzzword. *Feminist Theory,* 9(1), 67–85. https://doi.org/10.1177/1464700108086364.
- Decker, Marie C.; Wegner, Laila & Leicht-Scholten, Carmen. (2025). Procedural fairness in algorithmic decision-making: the role of public engagement. *Ethics and Information Technology,* 27(1). https://doi.org/10.1007/s10676-024-09811-4.
- Dwork, Cynthia; Hardt, Moritz; Pitassi, Toniann; Reingold, Omer & Zemel, Richard. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference.*
- Foulds, James; Islam, Rashidul; Keya, Kamrun N. & Pan, Shimei. (2018a). *Bayesian Modeling of Intersectional Fairness: The Variance of Bias*. https://arxiv.org/pdf/1811.07255.
- Foulds, James; Islam, Rashidul; Keya, Kamrun N. & Pan, Shimei. (2018b). *An Intersectional Definition of Fairness.* https://doi.org/10.48550/arXiv.1807.08362.
- Ghosh, Avijit; Genuit, Lea & Reagan, Mary. (2021). Characterizing Intersectional Group Fairness with Worst-Case Comparisons.
- Gikay, Asress A. (2020). The American Way — Until Machine Learning Algorithm Beats the Law? Algorithmic Consumer Credit Scoring in the EU and US. *SSRN Electronic Journal.* Vorab-Onlinepublikation. https://doi.org/10.2139/ssrn.3671488.
- Gohar, Usman & Cheng, Lu. (2023). A Survey on Intersectional Fairness in Machine Learning: Notions, Mitigation, and Challenges.
- Green, Ben. (2022). Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness. *Philosophy & Technology,* 35(4). https://doi.org/10.1007/s13347-022-00584-6.
- Hancock, Ange-Marie. (2007). When Multiplication Doesn't Equal Quick Addition: Examining Intersectionality as a Research Paradigm. *Perspectives on Politics,* 5(01). https://doi.org/10.1017/S1537592707070065.
- Hanna, Alex; Denton, Emily; Smart, Andrew & Smith-Loud, Jamila. (2020). Towards a critical race methodology in algorithmic fairness. In Association for Computing Machinery (Hrsg.), *Conference on Fairness, Accountability and Transparency (FAT* '20)* (S. 501–512). https://doi.org/10.1145/3351095.3372826.
- Hebert-Johnson, Ursula; Kim, Michael; Reingold, Omer & Rothblum, Guy. (2018). Multicalibration: Calibration for the (Computationally-Identifiable) Masses. In Jennifer Dy & Andreas Krause (Hrsg.), *Proceedings of Machine Learning Research, Proceedings of the 35th International Conference on Machine Learning* (S. 1939–1948). PMLR. https://proceedings.mlr.press/v80/hebert-johnson18a.html.

- Heinrichs, Bert. (2021). Discrimination in the age of artificial intelligence. *AI & SOCIETY.* Vorab-Onlinepublikation. https://doi.org/10.1007/s00146-021-01192-2.
- Henrich, Joseph; Heine, Steven J. & Norenzayan, Ara. (2010). The weirdest people in the world? *Behavioral and Brain Sciences,* 33(2–3), 61–83; discussion 83–135. https://doi.org/10.1017/s0140525x0999152x.
- Hoffmann, Anna L. (2019). Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society,* 22(7), 900–915. https://doi.org/10.1080/1369118X.2019.1573912.
- Islam, Rashidul; Keya, Kamrun N.; Pan, Shimei; Sarwate, Anand D. D. & Foulds, James R. R. (2023). Differential Fairness: An Intersectional Framework for Fair AI. *ENTROPY,* 25(4). https://doi.org/10.3390/e25040660.
- Kasirzadeh, Atoosa. (2022). Algorithmic Fairness and Structural Injustice: Insights from Feminist Political Philosophy. In *AIES '22: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. Association for Computing Machinery* (ACM) (S. 348–356). https://doi.org/10.1145/3514094.3534188.
- Kasy, Maximilian & Abebe, Rediet. (2021). Fairness, Equality, and Power in Algorithmic Decision-Making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (S. 576–586). ACM. https://doi.org/10.1145/3442188.3445919.
- Kearns, Michael; Neel, Seth; Roth, Aaron & Wu, Zhiwei S. (2018). Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning.*
- Kilbertus, Niki; Carulla, Mateo R.; Parascandolo, Giambattista; Hardt, Moritz; Janzing, Dominik & Schölkopf, Bernhard. (2017). Avoiding discrimination through causal reasoning. In Advances in Neural Information Processing Systems.
- Kim, Michael; Reingold, Omer & Rothblum, Guy. (2018). Fairness Through Computationally-Bounded Awareness. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi & Roman Garnett (Hrsg.), *Advances in Neural Information Processing Systems* (Vol. 31). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/c8dfece5cc68249206e4690fc4737a8d-Paper.pdf.
- Kong, Youjin. (2022). Are "Intersectionally Fair" AI Algorithms Really Fair to Women of Color? A Philosophical Analysis. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (S. 485–494). ACM. https://doi.org/10.1145/3531146.3533114.
- Krupiy, Tetyana. (2020). A vulnerability analysis: Theorising the impact of artificial intelligence decision-making processes on individuals, society and human diversity from a social justice perspective. *Computer Law & Security Review,* 38, 105429. https://doi.org/10.1016/j.clsr.2020.105429.
- Kusner, Matt J.; Loftus, Joshua; Russell, Chris & Silva, Ricardo. (2017). Counterfactual Fairness. *Advances in Neural Information Processing Systems,* 30, 4066–4076.
- Leavy, Susan; Siapera, Eugenia & O'Sullivan, Barry. (2021). Ethical Data Curation for AI: An Approach based on Feminist Epistemology and Critical Theories of Race. In *Conference on Artificial Intelligence, Ethics and Society (AIES),* Virtual Event, USA.
- Lee, Michelle S. A.; Floridi, Luciano & Singh, Jatinder. (2021). Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. *AI and Ethics,* 1(4), 529–544. https://doi.org/10.1007/s43681-021-00067-y.
- Mehrabi, Ninareh; Morstatter, Fred; Saxena, Nripsuta; Lerman, Kristina & Galstyan, Aram. (2019, 23. August). *A Survey on Bias and Fairness in Machine Learning.* http://arxiv.org/pdf/1908.09635v3.
- Mitchell, Shira; Potash, Eric; Barocas, Solon; D'Amour, Alexander & Lum, Kristian. (2021). Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application,* 8(1), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902.
- Nash, Jennifer C. (2017). Intersectionality and Its Discontents. *American Quarterly,* 69(1), 117–129. https://doi.org/10.1353/aq.2017.0006.
- Ovalle, Anaelia; Subramonian, Arjun; Gautam, Vagrant; Gee, Gilbert & Chang, Kai-Wei. (2023). Factoring the Matrix of Domination: A Critical Review and Reimagination of Intersectionality in AI Fairness. In Francesca Rossi, Sanmay Das, Jenny Davis, Kay Firth-Butterfield & Alex John (Hrsg.), *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (S. 496–511). ACM. https://doi.org/10.1145/3600211.3604705.
- Pessach, Dana & Shmueli, Erez. (2022). A Review on Fairness in Machine Learning. *ACM Computing Surveys,* 55(3), 1–44.
- Raghavan, Manish; Barocas, Solon; Kleinberg, Jon & Levy, Karen. (2020). Mitigating bias in algorithmic hiring. In Mireille Hildebrandt (Hrsg.), *ACM Digital Library, Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (S. 469–481). Asso-

ciation for Computing Machinery. https://doi.org/10.1145/3351095.3372828.
- Simson, Jan; Fabris, Alessandro & Kern, Christoph. (2024). Lazy Data Practices Harm Fairness Research. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (S. 642–659). ACM. https://doi.org/10.1145/3630106.3658931.
- Sundar, S. S. & Kim, Jinyoung. (2019). Machine Heuristic. In CHI '19, Proceedings of the 2019 *CHI Conference on Human Factors in Computing Systems* (S. 1–9). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300768.
- Walgenbach, Katharina. (2011). Intersektionalität als Analyseparadigma kultureller und sozialer Ungleichheiten. In Johannes Bilstein, Jutta Ecarius & Edwin Keiner (Hrsg.), *SpringerLink Bücher. Kulturelle Differenzen und Globalisierung: Herausforderungen für Erziehung und Bildung* (S. 113–130). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92859-3_7.
- Walgenbach, Katharina. (2012). Intersektionalität als Analyseperspektive heterogener Stadträume. In Elli Scambor & Carol Hagemann-White (Hrsg.), *Gender Studies. Die intersektionelle Stadt: Geschlechterforschung und Medienkunst an den Achsen der Ungleichheit* (S. 81–92). Transcript-Verlag https://doi.org/10.1515/transcript.9783839414156.81.
- Wang, A. J. (2018). Procedural Justice and Risk-Assessment Algorithms. *SSRN Electronic Journal.* Vorab-Onlinepublikation. https://doi.org/10.2139/ssrn.3170136.
- Yona, Gal & Rothblum, Guy. (2018). Probably Approximately Metric-Fair Learning. In Jennifer Dy & Andreas Krause (Hrsg.), *Proceedings of Machine Learning Research, Proceedings of the 35th International Conference on Machine Learning* (S. 5680–5688). PMLR. https://proceedings.mlr.press/v80/yona18a.html.
- Zezulka, Sebastian & Genin, Konstantin. (2024). From the Fair Distribution of Predictions to the Fair Distribution of Social Goods: Evaluating the Impact of Fair Machine Learning on Long-Term Unemployment. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (S. 1984–2006). ACM. https://doi.org/10.1145/3630106.3659020.
- Zimmermann, Annette & Lee-Stronach, Chad. (2022). Proceed with Caution. *Canadian Journal of Philosophy,* 52(1), 6–25. https://doi.org/10.1017/can.2021.17.

Kontakt und Information

Marie Decker, M.Sc.
RWTH Aachen University
Fakultät für Bauingenieurwesen
Gender und Diversity in den Ingenieurwissenschaften
Kackertstraße 9
52072 Aachen
marie.decker@gdi.rwth-aachen.de
www.gdi.rwth-aachen.de

Prof. Dr. phil. Carmen Leicht-Scholten
RWTH Aachen University
Fakultät für Bauingenieurwesen
Gender und Diversity in den Ingenieurwissenschaften
Kackertstraße 9
52072 Aachen
carmen.leicht@gdi.rwth-aachen.de
www.gdi.rwth-aachen.de

https://doi.org/10.17185/duepublico/82760

DOI: 10.17185/duepublico/82760
URN: urn:nbn:de:hbz:465-20241217-084303-6