Forens Psychiatr Psychol Kriminol
https://doi.org/10.1007/s11757-022-00741-9

ÜBERSICHT

# COMPAS: zu einer wegweisenden Debatte über algorithmische Risikobeurteilung

**Tim Räz**[1]

Eingegangen: 26. Juli 2022 / Angenommen: 21. September 2022


**Zusammenfassung**
„Correctional Offender Management Profiling for Alternative Sanctions“ (COMPAS) ist ein Risikobeurteilungsinstrument, das im Bereich der Strafjustiz in den USA eingesetzt wird. An COMPAS hat sich eine lebhafte Diskussion über Fairness entzündet, die bis heute andauert. Jedoch wurde diese Diskussion im deutschsprachigen Kontext bisher nicht stark rezipiert. In diesem Beitrag wird zuerst die Risikobeurteilung durch COMPAS systematisch dargestellt und diskutiert, wie COMPAS in den USA eingesetzt wird. Es wird dann auf drei wichtige Aspekte der Diskussion über COMPAS eingegangen, nämlich Fairness, Transparenz und Daten. Schließlich wird angedacht, welche Konsequenzen man aus der Diskussion für den deutschsprachigen Kontext ziehen kann.

**Schlüsselwörter** Fairness · Transparenz · Daten · Informatik · Strafjustiz

## COMPAS: on a pathbreaking debate on algorithmic risk assessment

**Abstract**
Correctional offender management profiling for alternative sanctions (COMPAS) is a risk assessment instrument currently used in the US criminal justice system. COMPAS has sparked a lively debate on fairness that still goes on today; however, this debate has barely been taken up so far in the German-speaking context. The present paper first provides a systematic account of risk assessment by COMPAS and discusses its current use in the USA. The paper then turns to three important aspects of the debate on COMPAS, viz. fairness, transparency, and data. Finally, the paper considers possible consequences of this discussion for the German-speaking context.

**Keywords** Fairness · Transparency · Data · Computer science · Criminal justice

## Einleitung

„Correctional Offender Management Profiling for Alternative Sanctions“ (COMPAS, „Profile zur Verwaltung von Straftäter*innen für alternative Sanktionen“, Übers. d. Autors) ist ein Computerprogramm, das in den USA zur Risikobeurteilung im Bereich der Strafjustiz eingesetzt wird. In den USA und in der Informatik wird COMPAS, insbesondere in Bezug auf Fairness, breit diskutiert. Erstaunlicherweise wird diese Diskussion in deutschsprachigen Foren zur Risikobeurteilung in der Strafjustiz nicht stark rezipiert; so ergibt etwa eine Volltextsuche in dieser Zeitschrift keinen Treffer für „COMPAS“ und andere relevante Schlagwörter. Dieser Artikel soll dazu beitragen, diese Lücke zu füllen. Dazu wird zuerst die Geschichte von COMPAS skizziert, die Funktionsweise dieses Risikobeurteilungsinstruments (RBI) zusammengefasst und dargestellt, wie COMPAS in den USA eingesetzt wird. Es folgt ein Abriss zur Debatte über COMPAS, die durch ProPublica, eine Non-Profit-Organisation für investigativen Journalismus, angestoßen wurde. Dabei werden die drei Aspekte Fairness, Transparenz und Datensätze genauer betrachtet. Schließlich wird überlegt, welche Lehren man aus der De-

COMPAS: „Correctional Offender Management Profiling for Alternative Sanctions“.

✉ Dr. Tim Räz
tim.raez@posteo.de

[1] Institut für Philosophie, Universität Bern, Bern, Schweiz

Published online: 05 October 2022

batte über COMPAS für den deutschsprachigen Raum und für die Strafjustiz ziehen kann.

## COMPAS im Überblick

COMPAS[1] wurde zuerst 1998 entwickelt und ist ein RBI in der aktuarischen Tradition.[2] Die Motivation für die Entwicklung von COMPAS war gemäß den Entwickler*innen, dass COMPAS im Vergleich zu älteren RBI mehr Faktoren berücksichtigt, die theoretisch relevant für das Rückfallrisiko sind.[3] Das wichtigste Ziel von COMPAS ist die Vorhersage des Rückfallrisikos von Straftäter*innen in Bezug auf allgemeine Straftaten und auf Gewaltstraftaten. Auch die Vorhersage von Risiken wie Nichterscheinen vor Gericht, von technischen Versäumnissen und von Disziplinarproblemen ist möglich. Als weitere Ziele von COMPAS werden die Planung und Überwachung der Rehabilitation im Strafvollzug genannt; auf die entsprechenden Funktionen wird hier nicht weiter eingegangen.

Die beiden Risikomodelle von COMPAS, eines für allgemeine Straftaten und eines für Gewaltstraftaten, sollen nun näher beschrieben werden. Zum Verständnis der Modelle sind drei Aspekte zentral: erstens die relevanten Faktoren (Inputvariablen), zweitens die Art, wie die Modelle diese Faktoren kombinieren, um zur Risikoprognose zu gelangen (innere Struktur), und drittens die Form der Prognose (Output-Variablen). Das Risikomodell für allgemeine Straftaten berücksichtigt Faktoren folgenden Typs: Alter, kriminelle Vorgeschichte, kriminelles Umfeld, Drogenmissbrauch und Indikatoren für jugendliche Delinquenz. Die Faktoren werden über einen Fragebogen mit zwischen 5 und über 10 Fragen/Faktortyp ermittelt. Das Modell für allgemeine Straftaten berücksichtigt insgesamt 70 Faktoren (Input-Variablen).[4] Die Struktur des Modells ist relativ einfach: Es handelt sich um ein lineares Modell, also um eine gewichtete Summe der Faktoren.[5] Das Modell prognostiziert das Risiko, innerhalb einer bestimmten Frist, z. B. innerhalb von zwei Jahren nach der Haftentlassung, eine weitere Straftat zu begehen und damit rückfällig zu werden. Die Vorhersage erfolgt in Form von Dezilen, also Zahlen von 1 bis 10, wobei ein Risiko von 1–4 als „tief", 5–7 als „mittel", und 8–10 als „hoch" interpretiert wird.[6] Die Anwendung des Modells dauert zwischen 10 und 60 Minuten pro Fall (Desmarais et al. 2018, S. 7). Das Risikomodell für Gewaltstraftaten berücksichtigt Faktoren folgenden Typs: bisherige Gewalttaten, Disziplinarprobleme, Probleme in Beruf und Bildung, Alter bei gegenwärtiger Verhaftung und Alter bei der ersten Verhaftung. Das Modell basiert ebenfalls auf einer Regression und macht Vorhersagen gleichen Typs wie das Modell für allgemeine Straftaten. Während die allgemeine Struktur der beiden Modelle bekannt und öffentlich zugänglich ist, werden die Details der Modelle, insbesondere die Gewichtung der Faktoren, geheim gehalten. COMPAS ist ein proprietäres Instrument der Firma Equivant/Northpointe (Maitland, FL, USA) und ist nicht öffentlich.

COMPAS wird gegenwärtig in fünf US-Bundesstaaten eingesetzt.[7] Gemäß der Rechteinhaberin Equivant/Northpointe wird COMPAS je nach Institution, in der COMPAS angewendet werden soll, gemäß den Bedürfnissen der Institution und dem Einsatzzweck angepasst und mit einer lokalen Normstichprobe kalibriert. Es gibt Versionen für verschiedene soziale Gruppen wie männliche und weibliche Straftäter*innen, Jugendliche sowie Versionen zum Einsatz während des Strafvollzugs und vor der Entlassung. COMPAS wurde in verschiedenen Untersuchungen empirisch überprüft (Desmarais et al. 2018). Gemäß dieser

[1] Dieser Abschnitt basiert auf der Darstellung in Brennan und Dieterich (2018), soweit nicht anders vermerkt. Man beachte, dass Tim Brennan und William Dieterich Entwickler*innen von COMPAS und Mitarbeiter*innen der Rechteinhaberin von COMPAS (Equivant, früher Northpointe) sind. Siehe auch Räz (im Druck) für eine Darstellung und kritische Würdigung von COMPAS.

[2] Aktuarische Risikobeurteilung ist eine Art von strukturierter Risikobeurteilung und kann von unstrukturierter Risikobeurteilung (Expertenurteilen) und auch von anderen strukturierten Methoden abgegrenzt werden; siehe etwa Desmarais et al. (2018). Gemäß Desmarais et al. ist strukturierte Risikobeurteilung prädiktiv erfolgreicher als unstrukturierte Risikobeurteilung.

[3] Der LSI-R ist gemäß den Entwickler*innen von COMPAS ein wichtiger Vorläufer von COMPAS; LSI-R war zum Zeitpunkt der Entwicklung von COMPAS ein RBI auf dem neuesten Stand, berücksichtigt aber gemäß den Entwickler*innen von COMPAS relevante Theorien der Kriminalität nicht.

[4] Die Angabe stammt aus Desmarais et al. (2018), einer Metaanalyse von RBI. Es ist erstaunlich schwierig, die genaue Anzahl Faktoren festzustellen, die in die beiden Risikomodelle einfließen. Im Rahmen der Recherche von ProPublica wurde ein Fragebogen zur Erhebung der genannten Faktoren publiziert (https://archive.epic.org/LibertyAtRiskReport.pdf), der insgesamt 137 Fragen, aufgeteilt auf 15 Faktortypen, umfasst. In der Darstellung der Entwickler*innen (Brennan und Dieterich 2018) sind keine genauen Angaben dazu zu finden. COMPAS erfasst mehr Faktoren als die meisten anderen Risikomodelle; so basiert etwa der LSI-R auf 54 Fragen (Desmarais et al. 2018).

[5] Das Modell wurde mittels logistischer Regression mit LASSO-Regularisierung konstruiert. Die LASSO-Regularisierung ist ein statistisches Standardverfahren zur Konstruktion linearer Modelle, die sicherstellt, dass irrelevante Variablen ausgeschlossen werden, siehe dazu etwa Hastie et al. (2009).

[6] Die Dezile werden aufgrund großer Normstichproben bestimmt. Wenn z. B. eine Person dem höchsten Dezil 10 zugewiesen wird, dann bedeutet das, dass diese Person ein prognostiziertes Rückfallrisiko hat wie die 10 % der Personen mit dem höchsten Rückfallrisiko der Normstichprobe.

[7] Siehe https://archive.epic.org/LibertyAtRiskReport.pdf.

Metaanalyse verschiedener RBI bewegt sich die prädiktive Validität des Modells für allgemeine Straftaten im Mittelfeld der untersuchten RBI.[8] Desmarais et al. monieren, dass es keine Überprüfung der Interrater-Reliabilität von COMPAS gibt.[9] Allerdings wird auch vermerkt, dass COMPAS eines der wenigen RBI ist, für die die prädiktive Validität in Bezug auf sozial relevante Gruppen (Gender, Ethnie[10]) untersucht wurde. An der Frage, wie prädiktive Modelle sich für verschiedene sozial relevante Gruppen verhalten sollten, hat sich ab 2016 eine Debatte entzündet, auf die im nächsten Abschnitt eingegangen wird.

## ProPublica vs. COMPAS: Fairness, Transparenz, Daten

In einem gewissen Sinn ist COMPAS nicht einfach nur ein RBI, sondern hat eine breitere, teilweise symbolische Bedeutung angenommen, in der Debatte über die Frage, ob und inwiefern Algorithmen fair sein können. Ausgangspunkt dieser Debatte ist eine Analyse von COMPAS durch ProPublica (Angwin et al. 2016). ProPublica beschaffte Daten von 7000 Personen, die in einem Bezirk in Florida verhaftet worden waren und deren Rückfallrisiko mit COMPAS vorhergesagt worden war. Diese Daten wurden mit Informationen über tatsächliche Rückfälle dieser Personen innerhalb von zwei Jahren nach der ersten Verhaftung abgeglichen. Ein Ergebnis dieser Untersuchung schlug besonders hohe Wellen: ProPublica verglich die Verteilung der Fehlerraten der beiden Risikomodelle separat für Schwarze Menschen und Weiße Menschen und stellte fest, dass das Modell für allgemeine Straftaten eine höhere Rate falsch-positiver Vorhersagen für Schwarze Menschen (45 %) ergab als für Weiße Menschen (23 %). Ebenso ergab das Modell eine höhere Rate falsch-negativer Vorhersagen für Weiße Menschen (48 %) als für Schwarze Menschen (28 %).[11] Diese Ergebnisse wurden von ProPublica dahingehend interpretiert, dass COMPAS Schwarze und Weiße Menschen ungleich behandelt und voreingenommen („biased“) gegenüber Schwarzen Menschen ist, weil Schwarze Menschen ohne Grund strenger beurteilt werden als Weiße Menschen.

Die Analyse von ProPublica schlug hohe Wellen in der Wissenschaft und in der Öffentlichkeit. Die Wissenschaft setzte sich einerseits mit COMPAS und der Analyse von ProPublica im engeren Sinn auseinander, die Intervention von ProPublica steht andererseits am Anfang einer breiteren wissenschaftlichen Debatte, die v. a. im neuen, interdisziplinär angelegten Feld der „fairness in machine learning“ (fair-ML) geführt wird und sich u. a. mit der Frage befasst, unter welchen (formalen) Bedingungen ein Modell wie COMPAS fair ist.[12] Die Dringlichkeit dieser Frage ergab sich u. a. aus kritischen Reaktionen auf die Analyse von ProPublica. So wiesen z. B. Flores et al. (2016) die Behauptung von ProPublica zurück, dass COMPAS gegenüber Schwarzen Menschen voreingenommen sei. Zwar stellten Flores et al. die gleiche, ungleiche Verteilung von Fehlerraten für die beiden Gruppen fest, betonten aber, dass eine Ungleichbehandlung nicht vorliege, wenn man die Kalibrierung nach Gruppen überprüfe, ein anderes Standardkriterium für Ungleichbehandlung aus der Forschung zu RBI.[13] Tatsächlich kommen die unterschiedlichen Beurteilungen von COMPAS dadurch zustande, dass ProPublica und etwa Flores et al. (2016) unterschiedliche Kriterien für Fairness angewandt haben. Bis heute gibt es keinen Konsens dazu, welches dieser Kriterien nun *tatsächlich* fair ist.[14] Die Frage nach der Kompatibilität und Bedeutung dieser und anderer Fairnesskriterien treibt die Debatte um fair-ML bis heute um.

Ein zweiter wichtiger Aspekt der Debatte ist die Intransparenz von COMPAS, ein Punkt, der etwa von Rudin et al. (2020) diskutiert wird. Rudin et al. kritisieren sowohl COMPAS als auch die Analyse von ProPublica.[15] Gemäß Rudin et al. ist COMPAS in zwei Hinsichten intransparent. Erstens seien die COMPAS-Risikomodelle unnötig komplex, da sie bis zu 137 Variablen berücksichtigten. Dies füh-

[8] Das Modell liegt in einem AUC-Bereich von 0,64–0,69 aufgrund von insgesamt drei Studien. AUC ist ein Maß für die Güte eines prädiktiven Modells, wobei ein AUC von 0,5 einer zufälligen (also schlechten) Vorhersagequalität entspricht und 1 einer perfekten Vorhersagequalität.

[9] Die Interrater-Reliabilität misst, zu welchem Grad verschiedene Rater, also Anwender eines RBI, ein und derselben Person die gleichen Faktorwerte zuordnen, die in die Berechnung der Prognose einfließen. Falls diese Zuverlässigkeit tief ist, kann dies dazu führen, dass verschiedene Rater die gleiche Person sehr unterschiedlich bewerten, was zu inkonsistenten Risikovorhersagen führen kann.

[10] Der Begriff „Ethnie“ wird in dieser Arbeit synonym zum englischen Begriff „race“ verwendet.

[11] Eine Vorhersage ist falsch-positiv, falls ein hohes Rückfallrisiko vorhergesagt wird, obwohl tatsächlich kein Rückfall stattfand, und falsch-negativ, falls ein tiefes Rückfallrisiko vorhergesagt wird, obwohl tatsächlich ein Rückfall stattfand. ProPublica stellte fest, dass die Fehlerraten des Modells für Gewaltstraftaten qualitativ ähnlich verteilt sind wie jene des Modells für allgemeine Straftaten.

[12] Ein wichtiges wissenschaftliches Forum für fair-ML ist die FAccT-Konferenz (FAccT steht für „fairness, accountability and transparency“), die seit 2018 durchgeführt wird. Während FAccT sich als fachübergreifend versteht, ist die Konferenz gemäß Form und Inhalt der Informatik zuzuordnen.

[13] Bei Kalibrierung für zwei Gruppen bildet man Dezile der COMPAS-Prognosen für die beiden Gruppen und vergleicht deren tatsächliche Rückfallwahrscheinlichkeiten in jedem Dezil. Tatsächlich sind diese für die beiden Gruppen bei COMPAS sehr ähnlich.

[14] Das Problem der Wahl verschiedener Fairnesskriterien wird dadurch verschärft, dass etwa die beiden erwähnten Kriterien unter milden Annahmen mathematisch inkompatibel sind, also notwendigerweise nicht gleichzeitig erfüllt werden können; Barocas et al. (2019) für einen Überblick zu diesen Fragen und Berk et al. (2018) für einen Überblick zur Fairness von RBI in der Strafjustiz.

[15] ProPublica wird u. a. dahingehend kritisiert, dass COMPAS tatsächlich nicht stark von der Variablen Ethnie abhänge, wenn man das Alter und die kriminelle Vergangenheit berücksichtige.

re zu einer höheren Fehleranfälligkeit, etwa aufgrund von Fehlern bei der Datenerfassung.[16] Zweitens sei COMPAS institutionell intransparent, weil die Rechteinhaberin Equivant/Northpointe die proprietären Modelle nicht veröffentliche. Schließlich behaupten Rudin et al., dass transparente, einfache und frei verfügbare Modelle mit nur zwei Variablen (Alter und Anzahl bisheriger Vergehen) eine ähnliche Vorhersagequalität liefern würden wie COMPAS. In Jackson und Mendoza (2020), einer Replik auf Rudin et al. von zwei Mitarbeiterinnen von Equivant/Northpointe, wird die Kritik an COMPAS zurückgewiesen. Zur unnötigen Komplexität merken Jackson und Mendoza an, dass beide Risikomodelle nicht alle 137 Variablen verwendeten; so basiere etwa das Modell für allgemeine Straftaten auf nicht mehr als 40 Variablen. Dieser Kritikpunkt von Jackson und Mendoza ist teilweise berechtigt. Auch die Struktur der (linearen) Risikomodelle ist nicht übermäßig komplex. Allerdings findet man zur verwendeten Anzahl der Faktoren verschiedene Angaben in unabhängigen Evaluationen der Risikomodelle (s. vorhergehenden Abschnitt). Den Vorwurf der institutionellen Intransparenz weisen Jackson und Mendoza zurück, da die Institutionen, die COMPAS anwendeten, vollen Zugang zur Struktur von COMPAS hätten. Weiter weisen Jackson und Mendoza auf mehrere unabhängige Studien hin, die COMPAS eine hohe prognostische Validität attestieren würden.[17] Die Antwort von Jackson und Mendoza auf die Kritik der institutionellen Intransparenz ist allerdings nicht überzeugend. Zum einen ist eine Offenlegung nur gegenüber anwendenden Institutionen ungenügend, weil damit unabhängige Wissenschaftler*innen und die Öffentlichkeit keinen hinreichenden Zugang zu den Risikomodellen haben. So ist es z. B. für die Nachvollziehbarkeit einer Risikovorhersage für betroffene Personen notwendig, dass Risikomodelle *inklusive einzelner Faktorgewichte* allgemein zugänglich sind. Auch für eine vollständige und unabhängige wissenschaftliche Evaluation müssten die Modelle publiziert werden.[18] Die Aussagekraft der von Jackson und Mendoza angeführten Studien zur prädiktiven Validität ist ebenfalls fraglich. Alle von Jackson und Mendoza zitierten Arbeiten sind technische Berichte von Institutionen, die COMPAS anwenden. Diese Evaluationen wurden nicht in begutachteten Zeitschriften publiziert und genügen damit einem wichtigen wissenschaftlichen Standard für Unabhängigkeit nicht. Weiter sind die Evaluationen von COMPAS nicht unabhängig im Sinn der Interessebindung.[19]

Ein dritter wichtiger Aspekt von COMPAS ist der Datensatz, der von ProPublica zur Evaluation verwendet wurde. Der COMPAS-Datensatz, von Angwin et al. (2016) aus verschiedenen Datenquellen kompiliert, wurde schon in ersten Reaktionen als inadäquat kritisiert. So wiesen etwa Flores et al. (2016) darauf hin, dass die Daten von Personen stammten, die noch kein Gerichtsverfahren durchlaufen hatten, während COMPAS eigentlich nur auf Personen angewendet werden sollte, bei denen die Haftstrafe zur Bewährung ausgesetzt worden war.[20] Trotz dieser bekannten Mängel wird der COMPAS-Datensatz bis heute als Benchmark in der Entwicklung fairer Algorithmen verwendet.[21] Diese Praxis ergibt sich teilweise aus der disziplinären Logik der Informatik. Die Informatik fokussiert stark auf die Entwicklung von Methoden, inklusive Fairnesskriterien, die sich optimieren und quantitativ vergleichen lassen. Um dies zu ermöglichen, werden oft dieselben Datensätze als Benchmark wiederverwendet. Dieses Vorgehen erlaubt einerseits einen schnellen methodischen Fortschritt, führt aber gleichzeitig zu einer Dekontextualisierung der Datensätze, sodass oft nicht mehr klar ist, inwiefern neue methodische Ergebnisse überhaupt auf den ursprünglichen Kontext eines Datensatzes anwendbar sind. Viele Probleme mit diesem Fokus auf Methodenoptimierung sind bekannt. So kann man den weiteren historischen Kontext aus den Augen verlieren, wie etwa unterschiedliche, historisch bedingte Rückfallraten sozial relevanter Gruppen. Solche historischen Fakten werden durch Datensätze reproduziert und durch Fairnesskriterien nicht unbedingt verändert. Auch vermeintliche Details wie in Datensätzen verwendete Codierungen und zur Erstellung von Datensätzen notwendige, aber nicht explizit gemachte Operationalisierungen können die Relevanz methodischer Ergebnisse stark beeinflussen (Bao et al. 2021).

## COMPAS und der deutschsprachige Raum

COMPAS ist ein RBI, das im Kontext des US-Justizsystems entwickelt wurde; die von ProPublica angestoßene

[16] Tatsächlich gibt es keine unabhängige Untersuchung zur Interrater-Reliabilität von COMPAS.

[17] Die prognostische Validität in den von Jackson und Mendoza zitierten Studien liegt deutlich höher als die in einer unabhängigen Metaanalyse festgestellten Werte (s. oben).

[18] Idealerweise würde der Quellcode der Modelle publiziert, wie es dem gegenwärtigen disziplinären Standard in der Informatik für reproduzierbare empirische Resultate entspricht. In Bezug auf die Datensätze ist die Lage etwas komplizierter, da hier der Datenschutz eine wichtige Rolle spielt.

[19] Als anwendende Institutionen stehen diese Institutionen unter einem Rechtfertigungsdruck für ihre Praxis der Risikobeurteilung und auch für den Einsatz von Mitteln, meist Steuergeldern, mit denen RBI wie COMPAS eingekauft werden.

[20] Möglicherweise haben in diesem Fall die Vollzugsbehörden in Florida die RBI nicht gemäß dem von Equivant/Northpointe vorgesehenen Verwendungszweck eingesetzt.

[21] Etwa in der Version von Friedler et al. (2019). Für eine detaillierte Diskussion des COMPAS-Datensatzes und seiner Bedeutung im Kontext der Strafjustiz: Bao et al. (2021). Im Sinn der Transparenz weist der Autor darauf hin, den COMPAS-Datensatz ebenfalls bereits in einer empirischen Arbeit als Benchmark verwendet zu haben.

Debatte ist ebenfalls stark auf den US-Kontext ausgerichtet. In diesem Abschnitt soll erörtert werden, welche Lehren man aus der Debatte über COMPAS für die Risikobeurteilung im deutschsprachigen Raum ziehen kann; exemplarisch soll dabei die deutschsprachige Schweiz betrachtet werden. Dies erfordert eine zweifache Übersetzung: zum einen eine Übertragung vom US-Kontext auf die Schweiz, zum anderen eine Übertragung von der Informatik auf die Risikobeurteilung in der Strafjustiz.

Zuerst soll überlegt werden, wie man die drei Aspekte Fairness, Transparenz und Datengrundlage im Kontext der Schweiz denken könnte. Die Debatte zur Fairness von COMPAS hat gezeigt, dass schwierige Güterabwägungen zur Ungleichbehandlung sozial relevanter Gruppen unausweichlich werden, sobald man Entscheidungen aufgrund von Risikovorhersagen fällt. Diese Güterabwägungen sind unabhängig davon, ob die Risikovorhersage von Expert*innen durchgeführt wird, oder ob es sich um strukturierte Risikobeurteilung handelt. Da in der Schweiz verschiedene RBI etwa für Gewaltstraftaten eingesetzt werden (Hahn 2016), werden solche Güterabwägungen faktisch bereits heute getroffen. Jedoch gibt es nur für einen Teil der gegenwärtig in der Schweiz eingesetzten RBI empirische Untersuchungen dazu, inwiefern verschiedene Gruppen ungleich behandelt werden.[22] In dieser Hinsicht hinkt die Schweiz der Entwicklung in den USA hinterher. Ungleichbehandlungen durch RBI in der Schweiz sind wohl eine Realität, aber sie haben wahrscheinlich eine andere Ausprägung als in den USA. In den USA fokussiert die Debatte stark auf die Diskriminierung Schwarzer Menschen. Während diese Form von rassistischer Diskriminierung durch RBI in der Schweiz ebenfalls ein Problem sein dürfte, sollten in diesem Kontext auch andere sozial relevante Gruppen berücksichtigt werden.[23] Die Frage, welche Gruppen hier berücksichtigt werden sollten, müsste diskutiert werden.

Die Debatte zur Intransparenz von COMPAS zeigt, dass für eine fundierte wissenschaftliche Auseinandersetzung eine vollständige Offenlegung von RBI nötig ist. Auch für die Erklärung von Einzelentscheiden ist eine solche Offenlegung unumgänglich. Im europäischen Kontext kommt hinzu, dass eine solche Offenlegung möglicherweise einem Rechtsanspruch im Rahmen der Datenschutz-Grundverordnung der EU entspricht (Goodman und Flaxman 2016). Auch das Problem der institutionellen Intransparenz ist nicht auf die USA und COMPAS beschränkt. So ist das im deutschsprachigen Raum eingesetzte Instrument FOTRES proprietär und nicht publiziert.[24] Für RBI müsste die institutionelle Transparenz auch im deutschsprachigen Raum eine Bedingung sein.[25] Außerdem wären Instrumente mit möglichst tiefer Komplexität vorzuziehen, um die Erklärbarkeit und leichte Anwendbarkeit sicherzustellen.[26] RBI müssen außerdem mithilfe lokaler Daten konstruiert und im lokalen Kontext getestet werden.

Dies führt zu den Daten, die zur Konstruktion und Evaluation von RBI benötigt werden. Viele der im vorhergehenden Abschnitt erwähnten Probleme mit Datensätzen im Kontext der Informatik sind im Kontext der forensischen Psychologie und Psychiatrie bekannt. So wurde etwa die „Hare Psychopathy Checklist-Revised“ (PCL-R) nach einer deutschen Adaptation mit Normdaten aus dem deutschsprachigen Raum neu validiert.[27] Allerdings kann man auch hier Unterschiede zum US-Kontext feststellen. Es fällt auf, dass für die Validierung der deutschsprachigen PCL-R ausschließlich männliche Probanden berücksichtigt wurden, die sich im Strafvollzug oder im Maßregelvollzug befanden. Der Einfluss von Gender oder Ethnie wurde nicht untersucht. Ein ähnliches Bild ergibt sich bei einer Validierungsstudie des bereits erwähnten RBI FOTRES (Rossegger et al. 2011; Räz im Druck). Am Beispiel FOTRES kann man einige der Schwierigkeiten antizipieren, falls man RBI für verschiedene Gruppen in der Schweiz untersuchen möchte. Beim Faktor Gender dürfte eine Schwierigkeit sein, dass wenige Daten für weibliche Straftäter*innen vorhanden sind, weil sich die Grundraten bezüglich Gender stark unterscheiden.[28] Beim Faktor Ethnie kann man ein Problem an der Stichprobe für die Validierung von FOTRES sehen. Personen ohne Schweizer Staatsangehörigkeit wurden für die Studie von

[22] Während es für den LSI-R, ein auch in der Schweiz eingesetztes Instrument, Untersuchungen bezüglich Gender und Ethnie gibt (mit gemischten Resultaten, allerdings im US-Kontext: Desmarais et al. 2018), gibt es für FOTRES, ein in der Schweiz entwickeltes und im deutschsprachigen Raum eingesetztes Instrument, keine unabhängige Evaluation der prädiktiven Validität (Habermeyer et al. 2020) und auch nicht bezüglich sozial relevanter Gruppen. Für eine kompakte Darstellung von FOTRES: Gonçalves et al. (2018); für einen systematischen Vergleich von FOTRES mit COMPAS: Räz (im Druck). Auch für unstrukturierte Risikobeurteilungen sind dem Autor keine empirischen Untersuchungen für den Kontext der Schweiz bekannt.

[23] So stellten Hangartner et al. (2021) in einer Untersuchung zur Diskriminierung von Stellensuchenden im Kontext der schweizerischen Arbeitsvermittlung fest, dass Menschen mit Herkunft aus Afrika, aber auch aus dem Mittleren Osten, Asien und Balkan besonders stark von Diskriminierung bei der Arbeitssuche betroffen sind.

[24] Mit „publiziert“ ist gemeint, dass der Quellcode des Risikomodells zugänglich gemacht wird, sodass Forschende und andere interessierte Parteien die Entscheidungslogik und -faktoren unabhängig überprüfen können; Räz (im Druck) für eine Diskussion von Transparenz.

[25] Bereits heute werden im deutschsprachigen Raum transparente und validierte RBI angewendet; Hahn (2016); Rettenberger (2016) für einen Überblick zu RBI für Gewaltstraftaten.

[26] Rudin et al. (2020) für Hinweise auf öffentlich verfügbare Modelle, die dafür infrage kommen.

[27] Während die PCL-R gemäß Hollerbach et al. (2018) ursprünglich als diagnostisches Instrument für psychopathische Merkmale entwickelt wurde, sei das Instrument auch prognostisch valide.

[28] Gemäß dem Bundesamt für Statistik waren 2019 6 % der Inhaftierten in der Schweiz weiblich.

Rossegger et al. (2011) nur berücksichtigt, wenn sie einen festen Aufenthaltsstatus hatten, da Personen ohne Schweizer Staatsangehörigkeit sonst am Ende des Strafvollzugs in ihr Herkunftsland „ausgeschafft“ werden. Da Ethnie mit Nationalität korreliert, erschwert dies die Feststellung von tatsächlichen Rückfallraten für verschiedene Gruppen und die Erhebung einer repräsentativen Stichprobe.

## Schlussfolgerung

Zum Schluss soll noch einmal überlegt werden, ob Instrumente wie COMPAS im deutschsprachigen Raum eingesetzt werden sollten, und welches die wichtigsten Erkenntnisse der Diskussion über COMPAS für den deutschsprachigen Raum sind.

Aufgrund der oben angestellten Überlegungen wird empfohlen, ein Instrument wie COMPAS nicht einzusetzen. Das wichtigste Argument gegen COMPAS ist der Mangel an Transparenz, insbesondere die Tatsache, dass COMPAS ein proprietäres Instrument ist. Ein solches Instrument genügt den Anforderungen an wissenschaftliche Transparenz und an die Nachvollziehbarkeit von Einzelfallentscheiden nicht. Das heißt nicht, dass alle RBI gleich problematisch sind. Grundsätzlich hat strukturierte Risikobeurteilung Vorteile, die nicht einfach ignoriert werden können, etwa die im Vergleich zu Expert*innenurteilen hohe prädiktive Kraft. Es sollte aber eine Wertedebatte darüber stattfinden, zu welchem Grad unsere Gesellschaft dazu bereit ist, Risiken von einzelnen Menschen aufgrund von allgemeinen, statistischen Mustern vorherzusagen; dies dürfte der Kern des Problems sein.

Zu den wichtigsten Fragen, welche die Diskussion über COMPAS aufwirft, gehört sicher, inwiefern ein RBI verschiedene sozial relevante Gruppen gleich oder verschieden beurteilen sollte, und was es genau heißt, verschiedene Gruppen fair zu behandeln. Es gibt bis heute auf diese Fragen keine eindeutigen Antworten. Trotzdem sollte eine Diskussion darüber geführt werden, insbesondere darüber, welche Gruppen im Kontext des deutschsprachigen Raums berücksichtigt werden sollten. Auch eine weitergehende interdisziplinäre Integration der Diskussion über RBI wäre wichtig. Während forensische Psychologie und Kriminologie teilweise von den Methoden der Informatik profitieren können, wäre es für die teilweise sehr abstrakten Debatten in der Informatik wünschenswert, wenn sie den empirischen Kontext von RBI in der Strafjustiz stärker berücksichtigen würden.

**Danksagung** Ich danke Claus Beisbart, Peer Briken und Corinna Hertweck für hilfreiche Kommentare zu einer früheren Fassung dieses Artikels.

**Funding** Die Arbeit an diesem Artikel wurde durch den Schweizerischen Nationalfonds gefördert (Projektnummer 197504). Die Open Access Finanzierung wurde durch die Universität Bern bereitgestellt.

**Interessenkonflikt** T. Räz gibt an, dass kein Interessenkonflikt besteht.



## Literatur

### Verwendete Literatur

Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. ProPublica

Bao M, Zhou A, Zottola S, Brubach B, Desmarais S, Horowitz A, Lum K, Venkatasubramanian S (2021) It's COMPASlicated: the messy relationship between RAI datasets and algorithmic fairness benchmarks (arXiv:2106.05498)

Barocas S, Hardt M, Narayanan A (2019) Fairness and machine learning. https://www.fairmlbook.org. Zugegriffen: 3 Okt 2022

Berk R, Heidari H, Jabbari S, Kearns M, Roth A (2018) Fairness in criminal justice risk assessments: the state of the art. Sociol Methods Res 50(1):3–44

Brennan T, Dieterich W (2018) Correctional Offender Management Profiles for Alternative Sanctions (COMPAS). In: Singh JP et al (Hrsg) Handbook of recidivism risk/needs assessment tools. Wiley-Blackwell, Hoboken

Desmarais SL, Johnson KL, Singh JP (2018) Performance of recidivism risk assessment instruments in U.S. Correctional settings. In: Singh JP et al (Hrsg) Handbook of recidivism risk/needs assessment tools. Wiley-Blackwell, Hoboken

Flores AW, Bechtel K, Lowenkamp CT (2016) False positives, false negatives, and false analyses: a rejoinder to machine bias: there's software used across the country to predict future criminals. And it's biased against blacks. Fed Probat 80(2):38–46

Friedler SA, Scheidegger C, Venkatasubramanian S, Choudhary S, Hamilton EP, Roth D (2019) A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the conference on fairness, accountability, and transparency, S 329–338

Gonçalves LC, Rossegger A, Endrass J (2018) Forensic Operationalized Therapy/Risk Evaluation System (FOTRES). In: Singh JP et al (Hrsg) Handbook of recidivism risk/needs assessment tools. Wiley-Blackwell, Hoboken

Goodman B, Flaxman S (2016) European Union regulations on algorithmic decision-making and a "right to explanation" (ArXiv:1606.08813v3)

Habermeyer E, Mokros A, Briken P (2020) "Die Relevanz eines kohärenten forensischen Beurteilungs- und Behandlungsprozesses": großer Wurf oder alter Wein in undichtem Schlauch? Forens Psychiatr Psychol Kriminol 14:212–219

Hahn S (2016) Violence risk assessment in Switzerland. In: Singh JP et al (Hrsg) International perspectives on violence risk assessment. Oxford University Press, Oxford, New York

Hangartner D, Kopp D, Siegenthaler M (2021) Monitoring hiring discrimination through online recruitment platforms. Nature 589(7843):572–576

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning, 2. Aufl. Springer series in statistics. Springer, New York

Hollerbach P, Mokros A, Nitschke J, Habermeyer E (2018) Hare Psychopathy Checklist-Revised. Deutschsprachige Normierung und Hinweise zur sachgerechten Anwendung. Forens Psychiatr Psychol Kriminol 12:186–191

Jackson E, Mendoza C (2020) Setting the record straight: what the COMPAS core risk and need assessment is and is not. Harv Data Sci Rev. https://doi.org/10.1162/99608f92.1b3dadaa

Räz T (2022) Understanding risk with FOTRES? AI and ethics (im Druck)

Rettenberger M (2016) The current status of sexual and violent recidivism and risk assessment research in Germany and Austria. In: Singh JP et al (Hrsg) International perspectives on violence risk assessment. Oxford University Press, Oxford, New York

Rossegger A, Laubacher A, Moskvitin K, Villmar T, Palermo GB, Endrass J (2011) Risk assessment instruments in repeat offending: the usefulness of FOTRES. Int J Offender Ther Comp Criminol 55(5):716–731

Rudin C, Wang C, Coker B (2020) The Age of secrecy and unfairness in recidivism prediction. Harv Data Sci Rev. https://doi.org/10.1162/99608f92.6ed64b30

## Weiterführende Literatur

Singh JP, Bjørkly S, Fazel S (Hrsg) (2016) International perspectives on violence risk assessment. Oxford University Press, Oxford, New York

Singh JP, Kroner DG, Wormith JS, Desmarais SL, Hamilton Z (Hrsg) (2018) Handbook of recidivism risk/needs assessment tools. Wiley-Blackwell, Hoboken