



Welche Art von Fairness macht KI-Systeme gerecht?

Eine Übersicht der verschiedenen Optionen. Und das Tool *Fairness Compass* für die optimale Auswahl.

Welche Art von Fairness macht KI-Systeme gerecht?

GETD | AI Research & Thought Leadership

1. Vorwort

Künstliche Intelligenz (KI) ist heutzutage in unserem Leben allgegenwärtig. Computer erkennen Muster und generieren aus großen Datenmengen Prognosemodelle, indem sie eine leistungsstarke KI-Technik verwenden, die allgemein als Machinelles Lernen (ML) bekannt ist. Banken und Versicherungen nutzen sie, um Kredit- oder Unfallrisiken vorherzusagen, Regierungen, um Sozialhilfe zu berechnen und Steuerbetrug zu erkennen, Supermärkte, um Profile ihrer Kunden zu erstellen, und Empfehlungssysteme, um Filme, Produkte, Jobs und Anzeigen vorzuschlagen. Wo in der Vergangenheit das Sammeln und Analysieren von Daten mit KI einem Unternehmen lediglich einen enormen Wettbewerbsvorteil verschaffte, ist das heute längst Standard, und Unternehmen, welche die neuen technischen Möglichkeiten nicht nutzen, werden wahrscheinlich keine Zukunft haben.

Diese durch KI beflügelte Welle des Fortschritts hat jedoch auch eine Kehrseite. So wurden kürzlich mehrere Fälle von Diskriminierung durch KI-Systeme aufgedeckt: KI-Sprachmodelle, die mit Nachrichtenartikeln trainiert worden waren, griffen gesellschaftliche Vorurteile auf; ein Tool zur Bewertung des Rückfallrisikos von Inhaftierten erwies sich als voreingenommen gegenüber Schwarzen; und Suchmaschinenergebnisse wurden durch rassistische und geschlechtsspezifische Stereotypen verfälscht. Und das sind nur einige Beispiele, die jüngst mediale Aufmerksamkeit erlangt haben. Entgegen erster Erwartungen haben sich die automatisch trainierten KI-Modelle nicht als unparteiische Richter erwiesen. Sind Diskriminierung und Voreingenommenheit im Kern der Vorhersagemodelle von automatisierten Entscheidungsverfahren einprogrammiert, führt der Einsatz solcher Systeme dazu, dass Vorurteile etabliert und verfestigt werden.

Als Reaktion auf dieses Problem wurden verschiedene Definitionen für Verzerrungen in Modellen und Daten vorgeschlagen, und Algorithmen wurden um Sicherheitsvorkehrungen erweitert, die Diskriminierung standardmäßig verhindern sollen. Festzulegen, was eine Verzerrung ausmacht und wie ein verzerrungsfreies Model aussieht, hat sich allerdings als nicht einfach erwiesen. Alle praxisnahen Modelle machen Fehler; es ist unvermeidlich, dass ein Modell, egal wie sorgfältig es trainiert wurde, hin und wieder falsche Entscheidungen trifft, wie

z.B. einer Person einen Kredit verweigert, die ihn eigentlich bekommen sollte.

Fairnessdefinitionen legen zugrunde, wie diese Fehler auf verschiedene Gruppen verteilt sind. Besteht etwa Ungleichheit zwischen den Geschlechtern oder zwischen verschiedenen Ethnien? Wie man Fehlerquoten unter den Gruppen idealerweise ausgleichen sollte ist allerdings keine triviale Frage, und die Antwort hängt von vielen Faktoren ab: dem Einsatzgebiet der Anwendung, Annahmen über die Daten, der sogenannten „Ground truth“, Vorgaben wie „Positive Diskriminierung“, ob Fairness auf Gruppen- oder auf individueller Ebene angestrebt wird usw. Wissenschaftliche Erkenntnisse, dass sich Fairnesskriterien, die aus ethischer Perspektive notwendig erscheinen, gegenseitig ausschließen, machen das Unterfangen nicht einfacher. Für Praktiker ist die Fülle an Methoden und Maßnahmen verwirrend und der Eindruck, dass jede Methode, egal wie sorgfältig sie angewendet wird, mindestens ein anderes Fairnesskriterium verletzt, mitunter entmutigend.

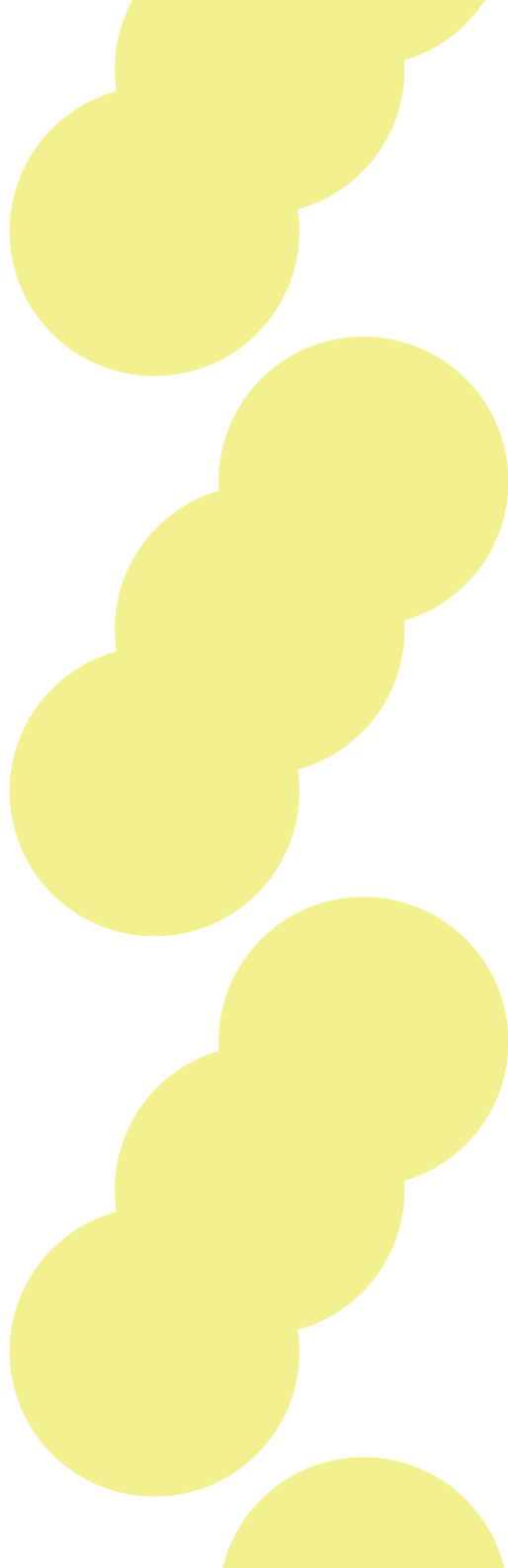
Dieser Kontext begründet die Motivation für ein Tool wie den *Fairness Compass*. Basierend auf die Erkenntnisse aus mehr als einem Jahrzehnt Forschung adressiert es das Problem des Informationsüberflusses, das die breite Anwendung von fairnessorientierter KI behindert, indem es Praktiker mithilfe eines schematischen Entscheidungsdiagramms unterstützt, jene Fairnessdefinition auszuwählen, die für eine konkrete Situation am besten geeignet ist. Das Auswahlverfahren wird so zu einem transparenten Prozess, der auf einer Reihe konkreter Fragen bezüglich der Art der Daten, Annahmen zu deren Verlässlichkeit, geltenden Fairness-Richtlinien, sowie der Frage, ob der Schwerpunkt eher auf der Spezifität oder der Sensitivität des Modells liegen sollte, basiert.

Der *Fairness Compass* ist ein lang ersehntes Tool für Maschinelles Lernen mit Fokus auf Fairness, und er wird Data Scientists und Praktikern den Weg zu einer faireren KI ebnen.



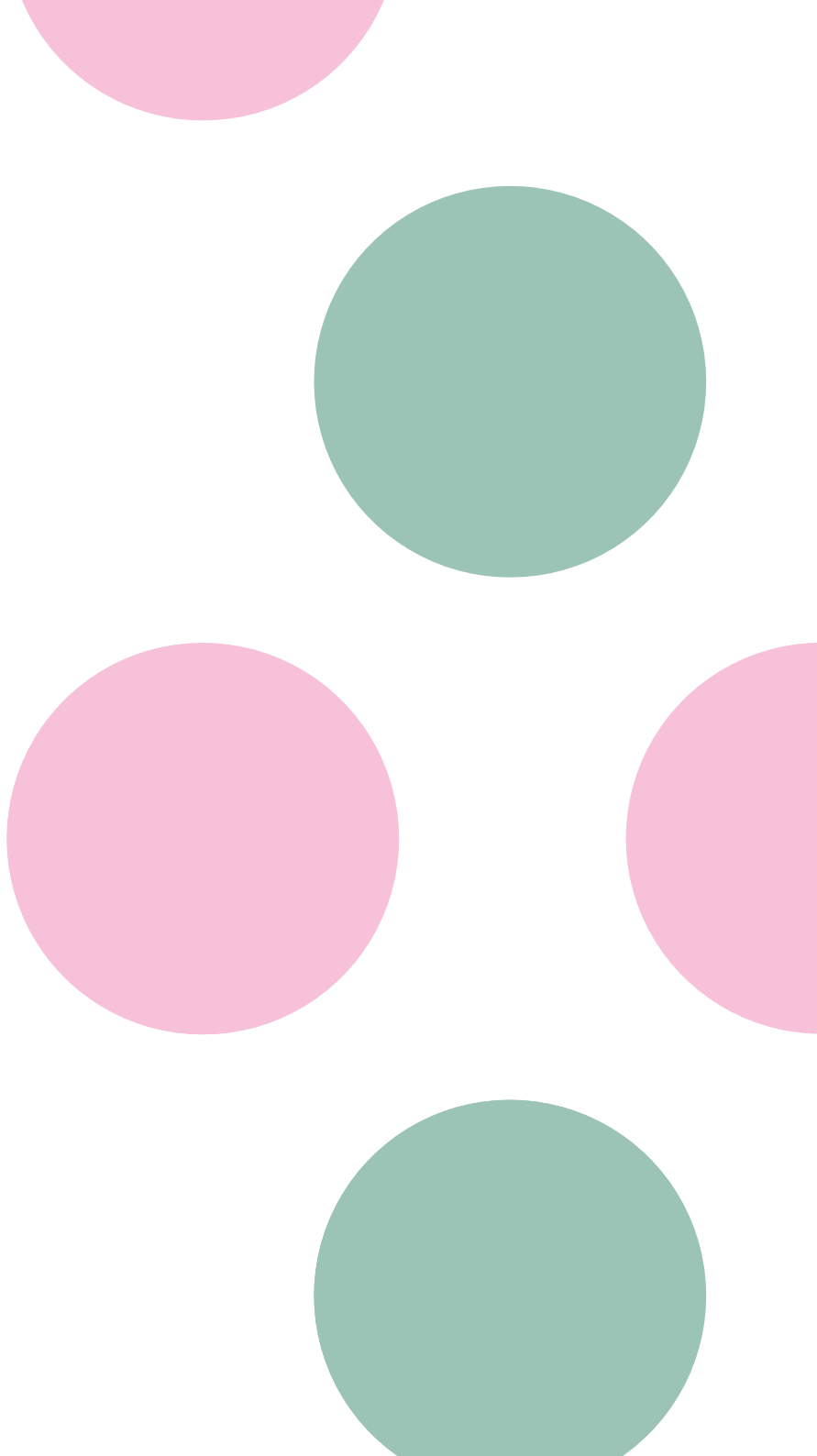
Toon CALDERS

Professor für Informatik an der Universität Antwerpen
und Wegbereiter in der Forschung zu KI-Fairness



Inhaltsverzeichnis

1.	Vorwort	4
2.	Einführung	9
3.	Grundlagen	12
3.1	Machine Learning	12
3.2	Statistische Kennzahlen	15
4.	Das Problem von Verzerrungen	21
5.	Verfügbare Fairnessdefinitionen	25
5.1	Unabhängigkeit	25
5.2	Suffizienz	29
5.3	Separierung	32
6.	Das Dilemma	36
7.	Der Fairness Compass	39
7.1	Anwendung	39
7.2	Entscheidungspunkte	40
7.3	Beispielanwendung	48
7.4	Weitere Entwicklung	48
8.	Schlussfolgerungen	49
9.	Literatur	50



2. Einführung

Im Laufe der letzten Jahre konnten bedeutende Durchbrüche im Bereich der Künstlichen Intelligenz (KI) erzielt werden. Ein großer Anteil an diesen Erfolgen ist auf Fortschritte beim Machinellen Lernen (ML) zurückzuführen, der Schlüsseltechnologie hinter kognitiven Systemen. Machinelles Lernen beschreibt die Fähigkeit einer Maschine, in großen Datensätzen Korrelationen zu erkennen. Aufgrund seiner Eigenschaft, riesige Mengen an Informationen in kurzer Zeit zu analysieren, können statistische Muster in Daten erkannt werden, die dem menschlichen Auge verborgen bleiben. Diese wiederum ermöglichen neuartige Erkenntnisse aus den Daten, welche die Datenanalyse und Modellprognosen optimieren helfen. Obwohl auch diese Ergebnisse nicht komplett fehlerfrei sind, so übertreffen sie doch die herkömmlichen

Ansätze, und schneiden oft sogar besser ab als menschliche Experten. Die Anwendungsgebiete für diese Technik sind vielfältig und umfassen medizinische Diagnosen, Studienplatzvergaben, Kreditbewilligungen, Rückfallprognosen, Rekrutierung, Onlinewerbung, Gesichtserkennung, Sprachübersetzung, Empfehlungssysteme, Betrugserkennung, Kreditkartenlimits, Preisgestaltung und Faktencheck in Sozialen Netzwerken.

Die große Abhängigkeit von den Daten birgt allerdings eine neue Herausforderung. Die Daten, die zum Training eines Algorithmus herangezogen werden, betrachtet man als *Ground Truth*. Das bedeutet, dass diese Daten während der Lernphase die vollumfängliche Realität abbilden, die das Prognosemodell anzunähern sucht. Sollten die Trainingsdaten auf irgendeine Weise unerwünschte Verzerrungen (*bias*) aufweisen, wird der trainierte Algorithmus diese widerspiegeln und sogar verstärken. Und weil die Logik von KI-Algorithmen für Menschen bislang nicht verständlich erklärbar ist, lassen sich diese Verzerrungen weder im Modell noch im Ergebnis ohne Weiteres als solche erkennen.

Von unerwünschten Verzerrungen können Teile der Gesellschaft betroffen sein, die sich über sensible Merkmale

wie zum Beispiel das Geschlecht, die ethnische Herkunft oder das Alter definieren. Demzufolge kann es vorkommen, dass Menschen aus diesen Gruppen von einem KI-System kategorisch benachteiligt werden. Systematische, ungleiche Behandlung von Individuen mit verschiedenen sensiblen Merkmalen wird als Diskriminierung betrachtet, und es herrscht breiter Konsens in unserer Gesellschaft, dass es unfair ist, wenn Ungleichbehandlung auf Grundlage persönlicher Eigenschaften stattfindet, auf welche der oder die Betroffene in der Regel keinen Einfluss hat. Entsprechend verbieten Antidiskriminierungsgesetze in vielen Ländern derartiges Handeln.

Bei statistischen Modellen, die mit traditionellen, deterministischen Algorithmen erzeugt werden, wird zur Vermeidung von Diskriminierung ein Verfahren eingesetzt, das als „Anti-Classification“ bekannt ist. Dieses Prinzip ist auch in der aktuellen Gesetzgebung verankert und sieht einfach vor, dass Datenparameter, die sensible Merkmale beschreiben, von der Verwendung ausgeschlossen sind. So darf zum Beispiel das Geschlecht einer Person in vielen Anwendungsfällen weder erhoben noch verwendet werden. Nun basiert Maschinelles Lernen allerdings auf „Big Data“, welche äußerst komplexe Korrelationen enthalten. Diese haben zur Folge, dass mitunter Parameter, welche unbedenklich scheinen und nicht als „sensibel“ eingestuft sind, indirekt sensible Informationen preisgeben können. Auf Grundlage dieser Verflechtungen konnte demonstriert werden, dass für KI-Systeme ein solcher Ansatz zur Bekämpfung von Diskriminierung ungenügend ist [1].

Im Entwicklungsprozess von ML-Modellen wurden zwei Hauptquellen für unerwünschte Verzerrungen festgestellt. Erstens kann es vorkommen, dass die Trainingsdaten fehlerhaft oder in bestimmter Hinsicht unzureichend repräsentativ sind. Derartige Mängel können die Ursache für Korrelationen in den Daten sein, die auf diese Weise in der Realität nicht zu finden sind. In so einem Fall erkennt der Algorithmus ein Muster, welches eigentlich keine Bedeutung hat. Zweitens ist es möglich, dass die Trainingsdaten zwar durchaus die Wirklichkeit abbilden, dieser Status Quo aber nicht der idealen Zielvorstellung entspricht. Wenn keine Korrektur erfolgt, reproduziert der Algorithmus den aktuellen Zustand und verfestigt so den bestehenden Missstand. Das Ziel besteht unter diesen Umständen darin, dass der finale Algorithmus die vorhandenen Verzerrungen ausgleicht.

In den letzten Jahren haben Forscherinnen und Forscher zahlreiche Methoden entwickelt, die diese Verzerrungen in den Daten, egal welchen Ursprungs, abschwächen, und KI-Systeme so fairer machen können. Das ist eine ermutigende Entwicklung, die hoffentlich das Vertrauen in KI stärkt und an deren Ende manche potentiell voreingenommene, menschliche Entscheidungen möglicherweise von unparteiischen, automatischen Entscheidungen ersetzt werden können. Neben der technischen Herausforderung, die Algorithmen oder die Daten anzupassen, gilt es allerdings eine ebenso wichtige, philosophische Frage zu klären: Welche Art von Fairness ist das Ziel? Fairness ist ein theoretisches Konzept von Gerechtigkeit, und es existieren verschiedene Definitionen, von denen manche untereinander in Konflikt stehen. Es gibt also keine universell anwendbare Form von Fairness, die allen Vorstellungen gleichermaßen genügt. Die optimale Definition hängt vielmehr vom konkreten Anwendungsfall ab und wird meistens von ethischen Grundsätzen und gesetzlichen Rahmenbedingungen bestimmt.

Dieses Dokument soll die Verantwortlichen für KI-Systemen unterstützen, die angestrebten ethischen Prinzipien anhand von Fragen und Beispielen festzulegen. Das vorgeschlagene Verfahren vereinfacht dabei nicht nur die Wahl der besten Fairnessdefinition für eine bestimmte Anwendung, sondern es macht diese Auswahl auch transparent und die implementierte Fairness für alle Beteiligten besser nachvollziehbar.

Die nachfolgenden Kapitel sind wie folgt strukturiert. Zunächst führen wir einige mathematische Grundlagen ein, die nützlich sind, um die Eigenschaften von Algorithmen des Maschinellen Lernens zu bewerten und zu vergleichen. Dann vertiefen wir das Problem von Verzerrungen in Daten. Anschließend präsentieren wir die am häufigsten verwendeten Definitionen von Fairness in der Forschung und erläutern die ethischen Prinzipien, die sie repräsentieren. Im folgenden Kapitel illustrieren wir beispielhaft, wie sich diese Fairnessdefinitionen mitunter gegenseitig widersprechen. Schließlich führen wir den *Fairness Compass* ein: unser praxisnahes Tool für KI-Entscheider, mit dem sich angestrebte ethische Standards in die passende Fairnessdefinition übersetzen lassen.




3. Grundlagen

Um die nachstehenden Kapitel besser verstehen zu können, vermitteln wir hier zunächst gewisse Grundkenntnisse zu Maschinellern und führen außerdem einige statistische Maße ein, die zur Prüfung und Bewertung der Systeme nützlich sind.

3.1 Machine Learning

Maschinelles Lernen unterscheidet sich von traditioneller Programmierung dadurch, dass die Logik des Algorithmus nicht auf hartkodierten Regeln beruht, die von einem Menschen so explizit festgelegt wurden, sondern vielmehr anhand von Beispielen erlernt wird: Tausenden, manchmal sogar Millionen Parameter werden ohne menschliches Eingreifen optimiert, um letztlich ein strukturelles Muster aus den Daten abzubilden. Das resultierende Prognosemodell ist dann im Stande, für neue Datensätze aus demselben Anwendungsbereich Vorhersagen mit hoher Präzision zu treffen.



Dieser Ansatz kann für zwei unterschiedliche Arten von Problemen eingesetzt werden: Für Klassifikation, wo die Aufgabe darin besteht, diskrete Klassen vorherzusagen, wie etwa Kategorien. Und für Regression, wo das Ziel ist, einen kontinuierlichen Wert zu prognostizieren, wie zum Beispiel einen Preis. Im vorliegenden Bericht gehen wir ausschließlich auf Klassifikationsprobleme ein, und der Einfachheit halber konzentrieren wir uns auf den binären Fall mit zwei Klassen: positiv (1) oder negativ (0). Als Ausgabewerte des Modells erwarten wir entweder eben jene Label 0 und 1; oder einen Score S , welcher die Wahrscheinlichkeit ausdrückt, dass eine Instanz positiv ist.

Um die hier vorgestellten Konzepte besser veranschaulichen zu können, verwenden wir ein fiktionales Szenario aus dem Bereich der Betrugserkennung bei Versicherungsfällen. Im Laufe der folgenden Kapitel werden wir uns immer wieder auf dieses Anwendungsbeispiel beziehen. Die Prüfung der Legitimität eines Versicherungsfalls ist unerlässlich um Missbrauch zu verhindern. Allerdings handelt es sich dabei für das Versicherungsunternehmen um einen aufwendigen und personalintensiven Vorgang. Zudem treten für manche Versicherungsarten viele Schadensfälle zeitgleich ein etwa durch Naturkatastrophen, die ganze Regionen betreffen. Für Versicherungsnehmer wiederum können genaue Kontrollen lästig sein, zum Beispiel wenn sie gebeten werden, weitere Fragen zu beantworten oder zusätzliche Dokumente nachzureichen. Beiden Parteien ist dabei an einer raschen Entscheidung gelegen: Die Kunden erwarten schnelle Abhilfe, und das Unternehmen versucht den Aufwand gering zu halten. Ein KI-System, das eine solche Aufgabe beschleunigt, könnte sich also als sehr nützlich erweisen. Konkret sollte es im Stande sein, rechtmäßige Versicherungsfälle sicher zu erkennen, um eine zeitnahe Auszahlung möglich zu machen. Potentiell betrügerische Fälle sollten ebenfalls zuverlässig entdeckt und für weitere Ermittlungen gekennzeichnet werden.

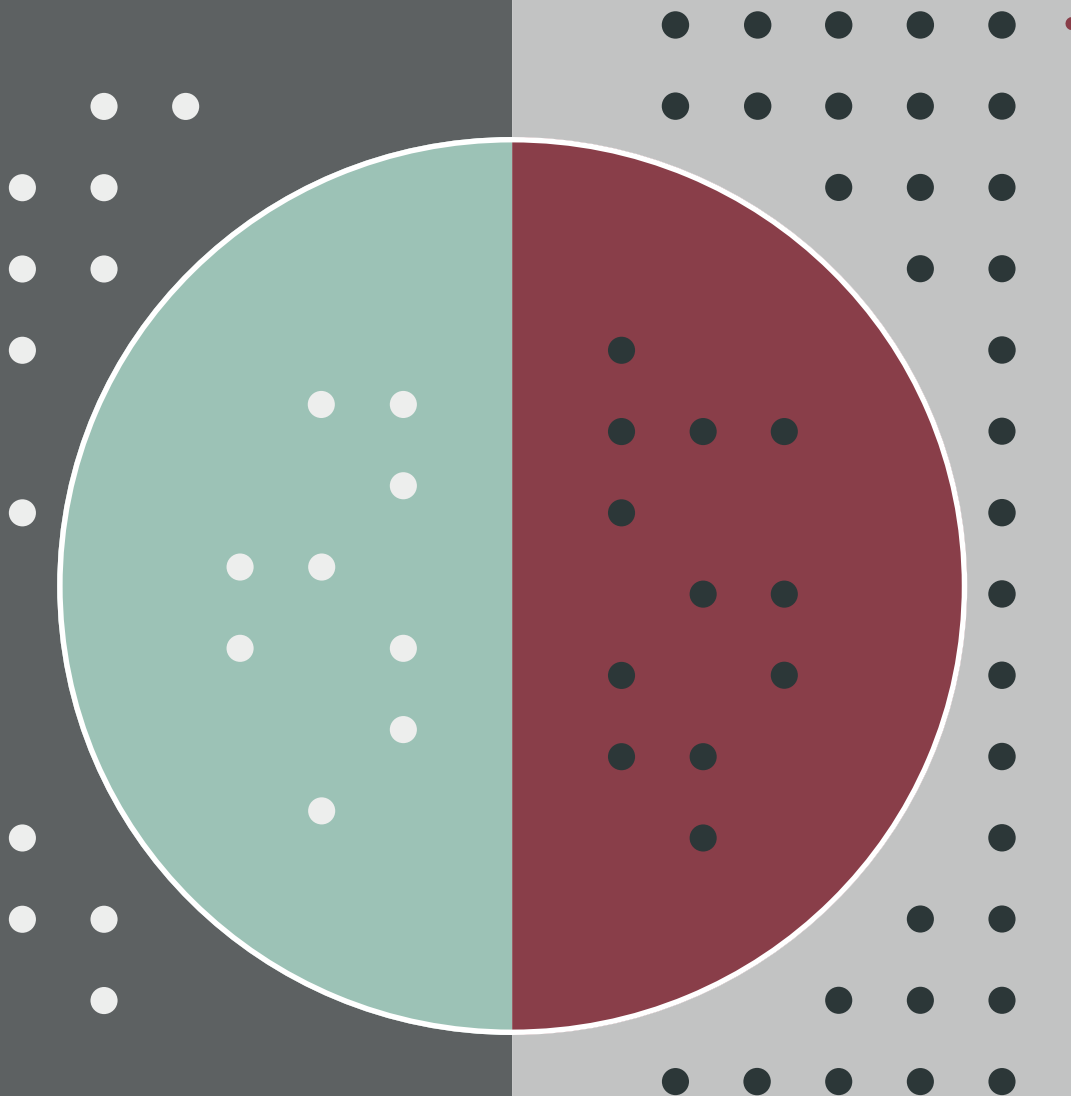


Abbildung 1: Grafische Darstellung der Ergebnisse aus **Tabelle 1**. Weiße Punkte repräsentieren unrechtmäßige Fälle (1), schwarze Punkte stehen für legitime Fälle (0). Der große weiße Kreis stellt das Vorhersagemodell dar.



3.2 Statistische Kennzahlen

Die sogenannte „Wahrheitsmatrix“ ist ein nützliches Mittel, um statistische Kennzahlen, die häufig zur Bewertung von Vorhersagemodellen herangezogen werden, darzustellen und zu berechnen. Die Zeilen der Matrix repräsentieren dabei die wahren Klassen, in unserem Fall 0 oder 1. Die Spalten beziehen sich auf die Vorhersagen des Modells. In den Zellen, wo die vorhergesagte Klasse mit dem tatsächlichen Ausgabewert übereinstimmt, stehen die Summen der jeweils korrekt klassifizierten Datensätze. Wo sich die Klassen unterscheiden, lag das Modell in seiner Vorhersage falsch und die Zellen enthalten die Summen der entsprechend fehlerhaft eingeordneten Fälle.

Auf abstrakter Ebene werden die Werte aus den Zellen üblicherweise mit den Begriffen aus [Tabelle 2](#) bezeichnet. Wenn wir die Daten aus unserem laufenden Beispiel in [Tabelle 1](#) als Grundlage nehmen, ergibt sich die Wahrheitsmatrix in [Tabelle 3](#). Wir stellen fest, dass das Vorhersagemodell in diesem Beispiel 9 Versicherungsfälle korrekt als

		Prognose	
		$\hat{Y}=1$	$\hat{Y}=0$
Wahr	$Y=1$	Richtig-positive Prognosen (<i>True positives TP</i>)	Falsch-negative Prognosen (<i>False negatives FN</i>)
	$Y=0$	Falsch-positive Prognosen (<i>False positives FP</i>)	Richtig-negative Prognosen (<i>True negatives TN</i>)

Tabelle 2: Schema einer Wahrheitsmatrix

		Prognose	
		$\hat{Y}=1$	$\hat{Y}=0$
Wahr	$Y=1$	9	12
	$Y=0$	12	30

Tabelle 3: Resultierende Wahrheitsmatrix aus den Beispieldaten in Tabelle 1

betrügerisch einstuft, und 30 ebenfalls richtig als legitim. Es klassifiziert jedoch auch 12 weitere Fälle als legitim, die eigentlich widerrechtlich sind, und 12 rechtmäßige Fälle zu Unrecht als Betrugsversuch.

Wenn wir uns erneut die grafische Darstellung in **Abbildung 1** vor Augen führen sehen wir nun, dass die farbig hinterlegten Segmente den unterschiedlichen Zellen in der Wahrheitsmatrix entsprechen: Falsch-negative Prognosen (dunkelgrau), richtig-positive Prognosen (grün), falsch-positive Prognosen (rot) und richtig-negative Prognosen (hellgrau).

Der Wahrheitsmatrix lassen sich zahlreiche statistische Kennzahlen entnehmen, die für eine Analyse des Modells interessant sind. Wir beschreiben diese im Einzelnen im nachfolgenden Text und stellen in **Tabelle 4** außerdem ihre Formeln und grafischen Darstellungen bereit.





























Tatsächlich positive Fälle	$P =$	$FN + TP$	
Tatsächlich negative Fälle	$N =$	$FP + TN$	
Basisrate	$BR =$	$\frac{P}{P + N}$	 
Positivrate	$PR =$	$\frac{TP + FP}{P + N}$	 
Negativrate	$NR =$	$\frac{TN + FN}{P + N}$	 
Korrektklassifikationsrate (Accuracy)	$ACC =$	$\frac{TP + TN}{P + N}$	 
Fehlklassifikationsrate (Misclassification rate)	$MR =$	$\frac{FN + FP}{P + N}$	 
Richtig-positiv-Rate (True positive rate)	$TPR =$	$\frac{TP}{P}$	 
Richtig-negativ-Rate (True negative rate)	$TNR =$	$\frac{TN}{N}$	 
Falsch-positiv-Rate	$FPR =$	$\frac{FP}{N}$	 
Falsch-negativ-Rate	$FNR =$	$\frac{FN}{P}$	 
Falscherkennungsrate (False discovery rate)	$FDR =$	$\frac{FP}{TP + FP}$	 
Positiver Vorhersagewert (Positive predictive value)	$PPV =$	$\frac{TP}{TP + FP}$	 
Falschausschlussrate (False omission rate)	$FOR =$	$\frac{FN}{TN + FN}$	 
Negativer Vorhersagewert (Negative predictive value)	$NPV =$	$\frac{TN}{TN + FN}$	 

Tabelle 4: Abgeleitete Formeln aus der Wahrheitsmatrix

Zunächst ermitteln wir die **tatsächlich positiven Fälle** im Datensatz. Diese Zahl ergibt sich aus den Summen der richtig-positiven und der falsch-negativen Prognosen. Letztere können auch als verfehlte positive Prognosen betrachtet werden. Entsprechend ergibt sich die Zahl der **tatsächlich negativen Fälle** aus der Summe der richtig-negativen und der falsch-positiven Prognosen, welche wiederum als verpasste negative Prognosen verstanden werden können. In unserem Beispiel entsprechen diese Kennzahlen den jeweiligen Summen der tatsächlich betrügerischen und der tatsächlich legitimen Versicherungsfälle.

Die (positive) **Basisrate**, manchmal auch als Prävalenzrate bezeichnet, steht für den Anteil der tatsächlich positiven Fälle, bezogen auf den kompletten Datensatz. Im Beispiel beschreibt diese Rate den wahren Anteil betrügerischer Fälle im Datensatz.

Die **Positivrate** wiederum ist der proportionale Anteil der positiven Bescheide, unabhängig davon, ob die Entscheidung richtig oder falsch war. Die **Negativrate** beschreibt umgekehrt die Rate der Negativebescheide, wieder ungeachtet dessen, ob die Entscheidung korrekt oder inkorrekt war. In unserem Beispiel entspricht die Positivrate der Rate jener Fälle, die als betrügerisch eingestuft wurden. Die Negativrate bezeichnet den Anteil der als legitim klassifizierten Fälle.

Die **Korrektklassifikationsrate** (*accuracy*) ist die Erfolgsquote, die den Anteil der korrekten Prognosen (positiv und negativ) von allen Entscheidungen bemisst. Im Gegenzug definiert die **Fehlklassifikationsrate** (*misclassification rate*) den Anteil der Fehlentscheidungen. In unserem Anwendungsbeispiel gibt die Korrektklassifikationsrate den Anteil der zurecht als legitim und der zurecht als betrügerisch eingeordneten Fälle wieder. Die Fehlklassifikationsrate bezieht sich auf die Fehlentscheidungen – der Anteil jener Instanzen also, bei der sich das Modell geirrt hat.

Die **Richtig-positiv-Rate** und die **Richtig-negativ-Rate** beschreiben die Proportionen der korrekt positiv bzw. korrekt negativ eingeordneten Instanzen, anteilig ihrer tatsächlichen Vorkommnisse. Im Beispiel steht die Richtig-positiv-Rate für den Anteil der tatsächlich betrügerischen Fälle, der vom Modell als solche erkannt wurde. Die Richtig-negativ-Rate beschreibt die Rate der tatsächlich legitimen Fälle, die erfolgreich als solche eingeordnet wurden.

In direktem Zusammenhang dazu beschreiben die **Falsch-positiv-Rate** und die **Falsch-negativ-Rate** die Fehlerquoten. Die Falsch-positiv-Rate bezeichnet den Anteil der eigentlich negativen Instanzen, der fälschlicherweise positiv klassifiziert wurde. Gleichmaßen steht die Falsch-negativ-Rate für die Rate der eigentlich positiven Fälle, die fälschlicherweise als negativ eingestuft wurden. Im Beispielszenario steht die Falsch-positiv-Rate für den Anteil der legitimen Fälle, welcher irrtümlich als betrügerisch klassifiziert wurde. Umgekehrt ist die Falsch-negativ-Rate der Anteil von den eigentlich betrügerischen Fällen, der vom System „übersehen“ und inkorrekt als legitim klassifiziert wurde.

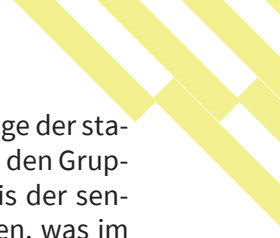
Die **Falscherkennungsrate** (*false discovery rate*) bezeichnet den Anteil der Fehlentscheidungen von allen positiv beschiedenen Fällen. Es geht also um den Anteil jener Instanzen, die zu Unrecht als positiv identifiziert bzw. entdeckt wurden. Andersherum beschreibt die **Falschauslassungsrate** (*false omission rate*) die Rate der fälschlicherweise negativ eingeordneten Fälle von allen Negativbescheiden. Diese eigentlich positiven Instanzen wurden ignoriert, d.h. sie wurden irrtümlich übergangen bzw. ausgelassen. Im Beispiel entspricht die Falscherkennungsrate der Fehlerrate von den Fällen, die als betrügerisch eingestuft wurden. Die Falschauslassungsrate bezeichnet hingegen die Fehlerrate von den als legitim vorhergesagten Versicherungsfälle – wie viele Fälle also irregulär als rechtmäßig eingestuft wurden.

Ähnlich, aber mit Schwerpunkt auf die korrekt klassifizierten Instanzen, beschreiben der **positive** und der **negative Vorhersagewert** den jeweiligen Anteil der positiven bzw. negativen Vorhersagen, der richtig war. Im Beispiel ist der positive Vorhersagewert der Anteil von jenen Prognosen, die einen Betrug vermuten, und damit richtig liegen. Der negative Vorhersagewert wiederum ist der Anteil der Fälle, die als legitim prognostiziert wurden, und das zurecht.


4. Das Problem von Verzerrungen

Bis jetzt haben wir für die statistische Analyse immer die Daten als Ganzes zugrunde gelegt, und nicht weiter berücksichtigt, dass der Datensatz aus sensiblen Untergruppen bestehen könnte. Die Entscheidungen von ML-Algorithmen betreffen jedoch oft Menschen. Daher ist es schon aufgrund der Beschaffenheit der Daten naheliegend, dass diese unterschiedliche demographische Gruppen umfassen, beispielsweise definiert durch das Geschlecht einer Person, deren ethnischen Hintergrund oder Konfession. Technisch wird die Zugehörigkeit zu einer solchen Gruppe meistens durch ein sensibles Attribut A im Datensatz festgehalten. Um ein Prognosemodell auf mögliche Verzerrungen zu prüfen, unterteilen wir die Ergebnisse mithilfe dieses sensiblen Attributs in verschiedene Datensätze und untersuchen deren statistische Merkmale auf Abweichungen. Unterschiedliche Kennwerte können Anzeichen von strukturellen Fehlern (*biases*) sein – verzerrte Prognosen also, die eine Ungleichbehandlung der sensiblen Untergruppen bedeutet.

Die Idee, Fairness auf Grundlage der Zugehörigkeit zu einer oder mehreren sensiblen Gruppen anzustreben, wird „Gruppenfairness“ genannt [2]. Dieser Ansatz findet sich auch in den Antidiskriminierungsgesetzen zahlreicher Gesetzgebungen wieder, mit verschiedenen Listen geschützter, sensibler Attribute [3, 4]. Alternativ gibt es in der Forschung ein weiteres Konzept namens „Individuelle Fairness“, das stattdessen eine Gleichbehandlung von Individuen angestrebt, die sich in sämtlichen Attributen ähneln – sensibel und nicht-sensibel [5]. Im Rahmen dieses Dokuments konzentrieren wir uns auf das Konzept der Gruppenfairness, und der Einfachheit halber gehen wir außerdem nur von zwei sensiblen Untergruppen aus. Folglich enthält unser Datensatz nur ein binäres, sensibles Attribut A , das die Werte 0 oder 1 annehmen kann, zum Beispiel um das Geschlecht zu kodieren.




Von unerwünschten Verzerrungen ist die Rede, wenn einige der statistischen Kennzahlen aus dem vorherigen Kapitel zwischen den Gruppen wesentlich abweichen. Es ist also notwendig, auf Basis der sensiblen Untergruppen separate Datenanalysen durchzuführen, was im Übrigen die Verfügbarkeit der sensiblen Attribute voraussetzt. Andernfalls lassen sich Probleme dieser Art nur schwer erkennen.



Wir untersuchen nun unser Anwendungsbeispiel zum Thema Betrugserkennung bei Versicherungsfällen auf Verzerrungen. Die Prognosedaten des trainierten Modells bleiben dabei unverändert, allerdings betrachten wir die Daten jetzt für zwei sensible Untergruppen, die durch das sensible Attribute A definiert sind. Zu diesem Zweck unterteilen wir die Daten für Männer ($A=0$) und Frauen ($A=1$). Anhand der unterschiedlichen Wahrheitsmatrizen für diese Untergruppen in [Tabelle 5](#) können wir jetzt deren statistische Eigenschaften inspizieren.

Wir stellen fest, dass die Basisraten (BR) für beide Gruppen identisch sind, was bedeutet, dass die Wahrscheinlichkeit, dass Männer und Frauen einen betrügerischen (oder einen legitimen) Schadensfall melden, gleich hoch ist. Die Richtig-negativ-Rate (TNR) liegt für Männer allerdings bei 0.79, während sie für Frauen 0.57 beträgt. Das bedeutet, dass 79% aller legitimen Fälle, die von Männern eingereicht wurden, korrekt als legitim eingeordnet wurden, wohingegen für Frauen das nur für 57% der Fälle des gleichen Typs gilt. Andererseits liegt die Falschausschlussrate (FOR) für Männer bei 24% und für Frauen bei 38%. Betrügerische Fälle, die von Frauen eingereicht werden, haben also eine höhere Chance unerkannt zu bleiben, als betrügerische Fälle von Männern.



		Prognose		
		$\hat{Y}=1$	$\hat{Y}=0$	
Wahr	A=0			BR=0.33
	Y=1	7	7	TPR=0.5
	Y=0	6	22	TNR=0.79
		FDR=0.46	FOR=0.24	
		PR=0.31	NR=0.69	

(a) Männer

		Prognose		
		$\hat{Y}=1$	$\hat{Y}=0$	
Wahr	A=1			BR=0.33
	Y=1	2	5	TPR=0.29
	Y=0	6	8	TNR=0.57
		FDR=0.75	FOR=0.38	
		PR=0.38	NR=0.62	

(b) Frauen

Tabelle 5: Unterschiedliche Wahrheitsmatrizen für sensible Untergruppen



5. Verfügbare Fairnessdefinitionen

Obwohl das Problem von Verzerrungen in KI-Systemen erst seit wenigen Jahren diskutiert wird, hat die Forschungsgemeinschaft bereits zahlreiche Lösungsvorschläge präsentiert. Dazu zählen etliche Fairnessdefinitionen, mit denen sich unerwünschte Verzerrungen in den Ergebnissen von Prognosemodellen, wie wir sie im vorherigen Kapitel beschrieben haben, messen lassen. Außerdem wurden verschiedene Methoden präsentiert, mit denen sich die jeweilige Art von Fairness erwirken lässt. Für weiterführende Informationen zu diesen Methoden verweisen wir auf entsprechende Übersichtsartikel als Ausgangspunkt [\[1, 2\]](#). In diesem Bericht konzentrieren wir uns auf die Fairnessdefinitionen und deren Auswirkungen auf die Ergebnisse in realen Anwendungsszenarien.

Im weiteren Verlauf stellen wir die am häufigsten verwendeten Fairnessdefinitionen für Gruppenfairness vor, und erklären deren Eigenschaften an Beispielen. Alle Definitionen lassen sich einem von drei statistischen Prinzipien zuordnen, die wir im Folgenden als Überkategorien verwenden: Die „bedingungslose“ Unabhängigkeit, sowie die bedingten Unabhängigkeiten Suffizienz und Separierung [\[6\]](#).

5.1 Unabhängigkeit

Statistisch betrachtet erfüllen Fairnessdefinitionen das Prinzip der Unabhängigkeit, wenn das sensible Attribut A von der Prognose \hat{Y} unbedingt unabhängig ist. Praktisch bedeutet das, dass auf alle Prognosen bezogen, der Anteil positiver und negativer Entscheidungen zwischen den sensiblen Gruppen proportional gleich ist. Auf individueller Ebene gilt dann für zwei Personen mit verschiedenen sensiblen Attributen, dass es für sie gleich wahrscheinlich ist, eine der beiden Klassen zugewiesen zu bekommen.




5.1.1 Demographic Parity

Das Ziel von Demographic Parity ist es, den sensiblen Untergruppen das vorteilhaftere Ergebnis in gleichen Raten zuzuweisen [5].

In unserem Beispiel ist das negative Ergebnis (=Klassifikation als legitimer Fall) das vorteilhaftere. Demographic Parity verlangt also negative Entscheidungen zu gleichen Raten für Männer und Frauen. Statistisch betrachtet müssen die Negativraten (NR) beider Untergruppen identisch sein. In der vorliegenden Verteilung (Tabelle 5) gilt allerdings für Männer $NR=0.42$ und für Frauen $NR=0.67$. Wir stellen also eine Abweichung von 25 Prozentpunkten für das vorteilhaftere Ergebnis zwischen den beiden sensiblen Untergruppen fest.

Die Wahrheitsmatrizen in Tabelle 6 enthalten eine mögliche Verteilung der Ergebnisse eines neuen Modells, das für Demographic Parity optimiert wurde. Die Zahl der negativen Prognosen für Männer ist gestiegen, die Matrix für Frauen bleibt unverändert. Beide Wahrheitsmatrizen weisen jetzt eine NR von 0.67 aus. Insofern wurde Demographic Parity erzielt. Allerdings ist es wenig überraschend, dass die Manipulation der Verteilung der Männer auch Änderungen der Richtig-positiv-Rate und der Richtig-negativ-Rate zur Folge hat.



		Prognose		
		$\hat{Y}=1$	$\hat{Y}=0$	
Wahr	A=0			BR=0.33
	Y=1	3	9	TPR=0.25
	Y=0	9	15	TNR=0.62
			FDR=0.75 FOR=0.38	
		PR=0.33 NR=0.67		

(a) Männer

		Prognose		
		$\hat{Y}=1$	$\hat{Y}=0$	
Wahr	A=1			BR=0.33
	Y=1	6	3	TPR=0.67
	Y=0	3	15	TNR=0.83
			FDR=0.33 FOR=0.17	
		PR=0.33 NR=0.67		

(b) Frauen

Tabelle 6: Optimiert für Demographic Parity

5.1.2 Conditional Statistical Parity

Diese Definition erweitert Demographic Parity durch eine Reihe vordefinierter Attribute, deren Einfluss auf die Prognose als legitim betrachtet wird. Das Ziel ist bei dieser Fairnessdefinition erreicht, wenn beide Untergruppen dieselben Chancen auf das vorteilhaftere Ergebnis haben, nachdem die Effekte der legitimen Kontrollvariablen herausgerechnet wurden [7].

In unserem Beispiel könnten vorherige Betrugsversuche eines Versicherungsnehmers die Wahrscheinlichkeit einer genaueren Untersuchung erhöhen. In diesem Fall wäre ein Attribut, das etwaige frühere Betrugsversuche dokumentiert, eine passende Kontrollvariable.

5.1.3 Equal Selection Parity

Während Demographic Parity gleiche Raten proportional zu den Gruppengrößen anstrebt, ist es das Ziel von Equal Selection Parity, dass jeder Untergruppe das bevorzugte Ergebnis in absoluten Zahlen gleich oft zugeteilt wird – unabhängig von den Gruppengrößen [8].

Im Beispiel der Betrugserkennung wäre diese Fairnessdefinition erfüllt, wenn für Männer und Frauen die gleiche Anzahl von Schadensfällen als legitim akzeptiert würde, selbst wenn eine Gruppe insgesamt mehr Fälle gemeldet hat, als die andere.



5.2 Suffizienz

Fairnesskonzepte erfüllen das Prinzip der Suffizienz, wenn das sensible Attribut A bedingt unabhängig von der wahren Klasse Y gegeben die Prognose \hat{Y} ist. Für die Prognosen einer jeden Klasse gilt also jeweils, dass A von Y unabhängig ist. In anderen Worten: Für alle positiven und alle negativen Vorhersagen ist der jeweilige Anteil korrekter Entscheidungen für beide sensiblen Untergruppen gleich. Auf individueller Basis lässt sich feststellen, dass Personen, die dieselbe Prognose erhalten haben, aber aus unterschiedlichen sensiblen Gruppen stammen, mit der gleichen Wahrscheinlichkeit der richtigen Klasse zugewiesen wurden.

5.2.1 Conditional Use Accuracy Equality

Entsprechend des Suffizienzkriteriums richtet sich diese Fairnessdefinition an der Prognose des Modells aus [9]. Statistisch betrachtet werden der positive Vorhersagewert (PPV) und der negative Vorhersagewert (NPV) für beide Gruppen angeglichen.

Im Zusammenhang mit unserem Beispiel heißt das, dass für alle Versicherungsfälle, die als betrügerisch klassifiziert wurden, diese Prognose für beide Untergruppen zu gleichen Teilen korrekt sein sollte. Und für die als legitim eingestuftten Fälle sollten diese Entscheidungen entsprechend für beide Gruppen gleichermaßen korrekt sein.

5.2.2 Predictive Parity

Bei Predictive Parity handelt es sich um eine abgeschwächte Form von Conditional Use Accuracy Equality, bei der die bedingte Unabhängigkeit nur den positiven Erwartungswert einschließt [10]. Folglich ist diese Fairnessdefinition bereits erfüllt, wenn bloß der positive Vorhersagewert (PPV) für beide Gruppen gleich ist.




5.2.3 Calibration

Calibration ist vergleichbar mit Conditional Use Accuracy Equality, wobei als bedingter Erwartungswert anstatt der binären Klassen ein Score S verwendet wird, der die Wahrscheinlichkeit einer Zuordnung zur positiven Klasse ausdrückt. Das Ziel ist wieder, für alle Untergruppen gleiche positive Vorhersagewerte (PPV) und gleiche negative Vorhersagewerte (NPV) zu erreichen [11]. Diese Form von Kalibrierung der Ergebnisse für beide Untergruppen kann mit identischen Wahrscheinlichkeiten einer korrekten (oder inkorrekten) Klassifizierung gleichgesetzt und entsprechend auch über eine Angleichung der Falscherkennungsrate (FDR) und der Falschausschlussrate (FOR) erreicht werden.

Im Rahmen unseres Beispiels hat eine Kalibrierung der Modellvorhersagen zur Folge, dass Männer und Frauen die gleichen Chancen haben, dass ihre eigentlich legitimen Fälle als betrügerisch eingestuft, oder dass unrechtmäßige Fälle fälschlicherweise akzeptiert werden.

Die beiden Wahrheitsmatrizen in **Tabelle 7** enthalten die kalibrierten Ergebnisse. Die Verteilung der Männer wurde angepasst, damit die FDR und die FOR den Werten der Frauen entspricht. Die Verteilung der Frauen wurde nicht verändert. Aufgrund der zugrundeliegenden identischen Basisraten in beiden Verteilungen hat diese Operation auch alle anderen statistischen Kennzahlen angeglichen.



		Prognose		
		$\hat{Y}=1$	$\hat{Y}=0$	
Wahr	Y=1	8	4	TPR=0.67
	Y=0	4	20	TNR=0.83
		FDR=0.33	FOR=0.17	
		PR=0.33	NR=0.67	

(a) Männer

		Prognose		
		$\hat{Y}=1$	$\hat{Y}=0$	
Wahr	Y=1	6	3	TPR=0.67
	Y=0	3	15	TNR=0.83
		FDR=0.33	FOR=0.17	
		PR=0.33	NR=0.67	

(b) Frauen

Tabelle 7: Optimiert für Calibration

5.3 Separierung

Fairnessdefinitionen erfüllen das Prinzip der Separierung, wenn das sensible Attribut A bedingt unabhängig vom Vorhersagewert \hat{Y} gegeben die wahre Klasse Y ist. Die Unabhängigkeit zwischen A und \hat{Y} ist also gegeben, wenn man die wahren Klassen separat betrachtet. Für sie gilt, dass die Anteile der korrekten Vorhersagen für beide sensible Untergruppen gleich sind. Individuell betrachtet garantiert diese Eigenschaft, dass zwei Personen, die zur gleichen Klasse gehören aber Mitglieder von verschiedenen sensiblen Untergruppen sind, die gleiche Chance auf eine richtige Einordnung haben.

5.3.1 Equalised Odds

Die Fairnessdefinition Equalised Odds erzielt Separierung: Für die wahren Klassen ist jeweils Unabhängigkeit gewährleistet, indem die Richtig-positiv und Richtig-negativ-Raten für die sensible Untergruppen gleich sind [12]. Die Überlegung hinter diesem Konzept ist, dass die Chancen auf eine korrekte Klassifikation für alle Personen gleich sein sollten.

Auf unser wiederkehrendes Beispiel bezogen bedeutet Equalised Odds, dass für Männer und Frauen die Chancen gleich sind, dass ihre Versicherungsfälle zurecht als legitim oder betrügerisch eingeordnet werden; das Prognosemodell sollte nicht für eine Untergruppe mehr oder weniger präzise funktionieren, als für die andere.

Tabelle 8 zeigt ein mögliches Ergebnis für unsere Beispielverteilungen, das Equalised Odds erfüllt. Die Prognosen für Männer wurden angepasst, damit die Richtig-positiv-Rate (TPR) und die Richtig-negativ-Rate (TNR) denen der Frauen entsprechen. Da beide Gruppen die gleichen Basisraten haben, hat dieser Vorgang zur Folge, dass sich auch die übrigen statistischen Kennzahlen angeglichen haben.

		Prognose		
		$\hat{Y}=1$	$\hat{Y}=0$	
Wahr	A=0			BR=0.33
	Y=1	8	4	TPR=0.67
	Y=0	4	20	TNR=0.83
			FDR=0.33	FOR=0.17
		PR=0.33	NR=0.67	

(a) Männer

		Prognose		
		$\hat{Y}=1$	$\hat{Y}=0$	
Wahr	A=1			BR=0.33
	Y=1	6	3	TPR=0.67
	Y=0	3	15	TNR=0.83
			FDR=0.33	FOR=0.17
		PR=0.33	NR=0.67	

(b) Frauen

Tabelle 8: Optimiert für Equalised Odds

5.3.2 Equalised Opportunities

Bei komplexen Daten aus echten Anwendungen kann sich das Erzielen von Equalised Odds als schwierig erweisen. Daher wurde die Fairnessdefinition Equalised Opportunities als praktikablere Alternative vorgeschlagen [12]. In dieser abgeschwächten Version von Equalised Odds muss lediglich die Fehlerrate für die positive Prognose identisch sein.

Im Beispielkontext ist Equalised Opportunities erfüllt, wenn tatsächlich betrügerische Fälle von Männern und Frauen zu gleichen Raten abgelehnt werden. Für legitime Fälle darf die Rate der akzeptierten Fälle zwischen den beiden Gruppen abweichen.

5.3.3 Predictive Equality

Eine weitere Abwandlung von Equalised Odds ist Predictive Equality. Hierbei müssen sich nur die Fehlerraten für die negative Prognose entsprechen [13].

In unserem Beispiel ist Predictive Equality erfüllt, wenn Männer und Frauen erwarten können, dass ihre legitimen Versicherungsfälle zu gleichen Raten genehmigt werden. Die Fehlerraten für betrügerische Fälle können indessen bei dieser Fairnessdefinition für die beiden Untergruppen abweichen.

Wie zuvor bleibt die Wahrheitsmatrix für Frauen in **Tabelle 9** unverändert. Für Männer wurde die Verteilung dahingehend angepasst, dass die Falsch-positiv Rate (TNR) mit der der Frauen übereinstimmt. Die Fehlerraten für das unvorteilhaftere Ergebnis, dass ein Fall nämlich als betrügerisch eingestuft wird, können weiterhin für die beiden Geschlechter abweichen.

5.3.4 Balance

Alle vorherigen Fairnessdefinitionen, die sich auf das Prinzip der Separierung stützen, haben Prognosen in Form binärer Ausgabewerte verwendet. Bei der Definition Balance werden stattdessen die Wahrscheinlichkeiten zugrunde gelegt, und deren Durchschnittswerte je Klasse für beide Gruppen verglichen. Dieser Ansatz soll verhindern, dass die Prognosen für eine Gruppe systematisch niedriger ausfallen, was im binären Fall eventuell nicht weiter auffallen würde. Stattdessen

strebt diese Fairnessdefinition ausgeglichene Ergebnisse auf Basis der gemittelten Wahrscheinlichkeitswerte je Gruppe an. Abhängig vom Anwendungsfall ist es möglich, eine entsprechende Balance für die positive oder für die negative Klasse anzustreben [14].

		Prognose		
		$\hat{Y}=1$	$\hat{Y}=0$	
Wahr	A=0			BR=0.33
	Y=1	9	3	TPR=0.75
	Y=0	4	20	TNR=0.83
		FDR=0.31	FOR=0.13	
		PR=0.36	NR=0.64	

(a) Männer

		Prognose		
		$\hat{Y}=1$	$\hat{Y}=0$	
Wahr	A=1			BR=0.33
	Y=1	6	3	TPR=0.67
	Y=0	3	15	TNR=0.83
		FDR=0.33	FOR=0.17	
		PR=0.33	NR=0.67	

(b) Frauen

Tabelle 9: Optimiert für Predictive Equality

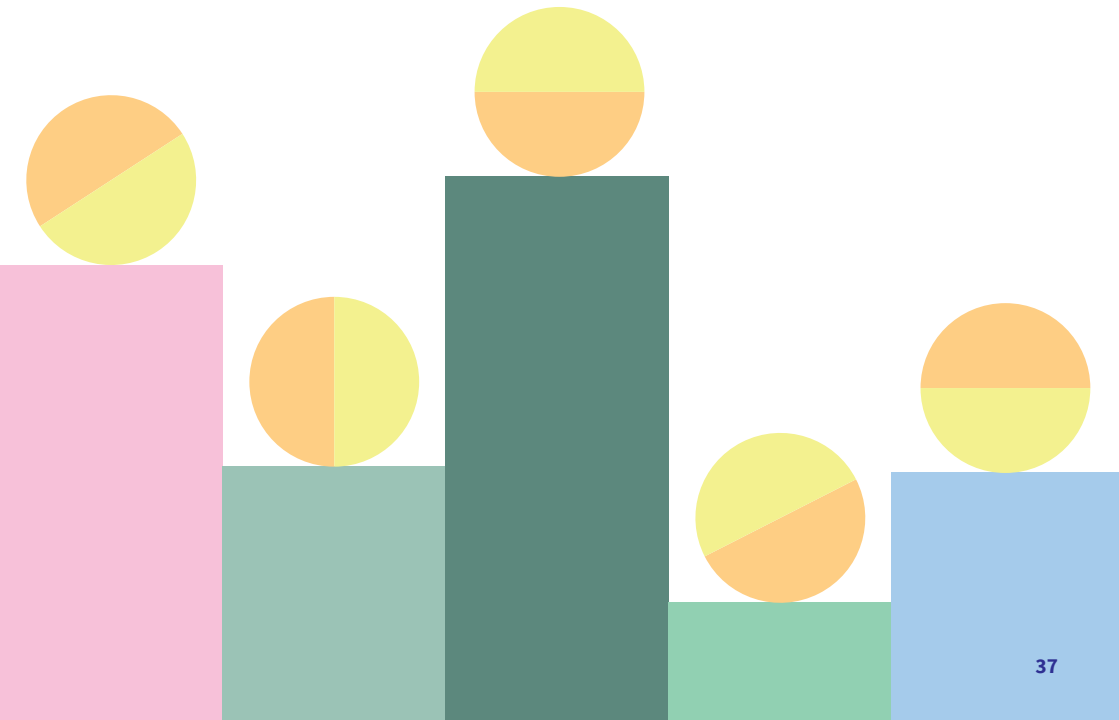
6. Das Dilemma

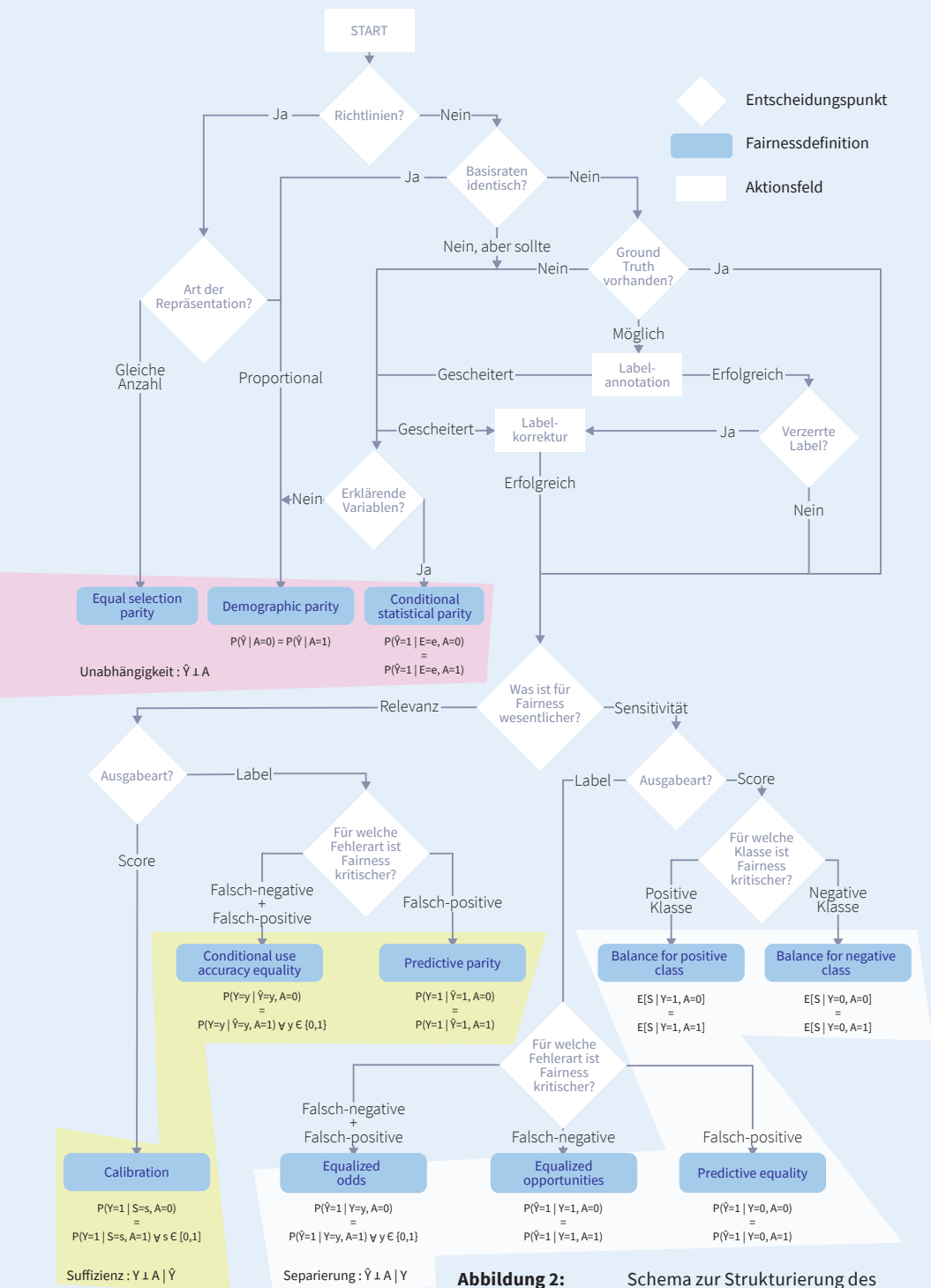
Angesichts all dieser unterschiedlichen Fairnessdefinitionen ist der Wunsch naheliegend, eine Art von „kompletter Fairness“ zu finden – eine ultimative Lösung also, die allen Typen von Fairness gleichermaßen gerecht wird. Manche der Fairnessformeln hängen jedoch über gemeinsame Variablen zusammen, und für einige von ihnen wurde mathematisch nachgewiesen, dass es zumindest unter Praxisbedingungen unmöglich ist, sie parallel zu optimieren [14, 15, 1]. Stattdessen bringen Verbesserungen für die eine Fairnessdefinition immer Verschlechterungen für eine andere mit sich. Wenn man die Zusammenhänge zwischen den Fairnessdefinitionen und die konditionalen Beziehungen innerhalb der Wahrheitsmatrix betrachtet, ist der Grund hierfür leicht nachvollziehbar: Die Formeln verwenden mitunter dieselben Zellwerte, und die Zellen selbst stehen in Beziehung zueinander (z.B. sind die Summen der Zeilen, welche die tatsächlichen Vorkommnisse für die jeweilige Klasse darstellen, fix).

In der Öffentlichkeit ist vor allem der Kompromiss zwischen kalibrierten Prognosen (Calibration) sowie gleichen Falsch-positiv und Falsch-negativ-Raten (gleichzusetzen mit Equalised Odds) diskutiert worden. Die Debatte wurde durch einen ML-Algorithmus namens „Correctional Offender Management Profiling for Alternative Sanctions“, kurz COMPAS, ausgelöst. Dieser war von der Firma Northpointe, Inc. entwickelt worden, und sein Zweck bestand darin, einen unabhängigen, datengestützten „Risikoscore“ für verschiedene Formen von Rückfallkriminalität zu ermitteln. Algorithmen dieser Art werden in den USA in der Strafjustiz verwendet, um den Richter oder die Richterin bei bestimmten Entscheidungen wie der Gewährung von Kaution oder Bewährung zu unterstützen. Der Score hat rein informativen Charakter, die finale Entscheidung liegt weiterhin beim Mensch. Im Mai 2016 hat ProPublica, eine Plattform für investigativen Journalismus, mit einem Artikel für Aufsehen gesorgt, der dem COMPAS-Algorithmus rassistische Entscheidungen vorwirft [16]. Das Hauptargument des Berichts stützt sich auf eine Datenanalyse, die Verzerrungen bei den Prognosen feststellt. Insbesondere fiel die Falsch-positiv-Rate für Schwarze deutlich höher aus, als für Weiße. Konkret bedeutet das,

dass Schwarzen überproportional häufig Prognosen ausgestellt werden, die zu Unrecht eine erhöhte Rückfälligkeit suggerieren. Northpointe widersprach dem Vorwurf der Diskriminierung mit dem Argument, dass ihr Algorithmus durchaus faire Entscheidungen trafe, indem er Predictive Parity für beide Gruppen erziele: Der Risikoscore gibt also die Wahrscheinlichkeit eines Rückfalls wieder, und das mit gleicher Zuverlässigkeit für beide Gruppen [17].

Objektiv betrachtet kann festgehalten werden, dass die Argumente beider Seiten richtig und berechtigt sind. Die hitzige Debatte allerdings hat deutlich gemacht, dass es für eine KI-Anwendung unumgänglich ist, die angestrebte Art von Fairness vorab festzulegen und zu kommunizieren. Die Wahl bedarf in der Regel Abwägungen und Kompromisse. Im vorliegenden Fall zum Beispiel könnten Calibration und Equalised Odds theoretisch nur dann gemeinsam erreicht werden, wenn eine der beiden folgenden Konditionen eintritt: Entweder, wenn die beiden Basisraten der Untergruppen exakt identisch sind, oder wenn sich die Klassen perfekt separieren lassen. Letzteres würde nämlich das Erstellen eines idealen, fehlerfreien Prognosemodells möglich machen. Leider sind beide Voraussetzungen unter realen Bedingungen sehr unwahrscheinlich.





7. Der Fairness Compass

Aufgrund der im vorherigen Kapitel beschriebenen Einschränkungen sollte beim Einsatz von Künstlicher Intelligenz für jeden Anwendungsfall vorab sorgfältig die passende Definition von Fairness ausgewählt werden. Um KI-Entscheider in dieser Aufgabe zu unterstützen, haben wir den *Fairness Compass* entwickelt: ein Schema in Form eines Entscheidungsbaums, das den Auswahlprozess systematisch vereinfacht. Dabei hilft die Struktur, die ethischen Grundprinzipien für eine Anwendung festzulegen und die jeweiligen Argumente zu dokumentieren. Im Ergebnis formalisiert dieses Tool den Entscheidungsprozess und ermöglicht es so, die Wahl der implementierten Art von Fairness beispielsweise dem Verbraucher detailliert zu begründen.

In diesem Kapitel beschreiben wir zunächst die allgemeine Anwendung des Tools und erläutern dann die Knotenpunkte des Entscheidungsbaums im Einzelnen. Schließlich erklären wir einige technische Einzelheiten und führen aus, wie sich dieses Projekt weiterentwickeln könnte.

7.1 Anwendung

In erster Linie besteht das Tool aus dem Entscheidungsbaum in **Abbildung 2**, der den Auswahlprozess beschreibt. Das Diagramm enthält drei verschiedene Arten von Symbolen: die Rauten stellen die Entscheidungspunkte dar, die weißen Boxen sind Aktionsfelder, und die grauen Boxen mit runden Ecken symbolisieren die jeweiligen Fairnessdefinitionen. Die Pfeile, welche die Symbole verbinden, repräsentieren die Wahlmöglichkeiten. Um die Nutzbarkeit zu vereinfachen ist das Schema auch als interaktives **Online Tool**¹ verfügbar. In dieser Version kann man den Entscheidungsbaum leichter erkunden, da sich Tooltips mit weiterführenden Informationen, Beispielen und Referenzen einblenden lassen. Das interaktive Tool bietet sich außerdem dazu an, den Entscheidungsprozess für einen bestimmten Anwendungsfall

1 <https://axa-rev-research.github.io/fairness-compass.html>

hervorzuheben, und die Begründung für jede Entscheidung in Form von Tooltips zu hinterlegen. Auf diese Weise nützt das Tool nicht nur KI-Verantwortlichen bei der Entscheidungsfindung, sondern es eignet sich auch als Mittel, dem Verbraucher diese Entscheidung zu erklären. KI-Systeme fair zu gestalten ist ein vielschichtiges Thema und bedarf kontextbezogener Lösungen. Wir sind daher überzeugt, dass die breitere Öffentlichkeit in die Details einbezogen werden sollte, um das Vertrauen in KI-gestützte Anwendungen langfristig zu stärken.

7.2 Entscheidungspunkte

Im Folgenden präsentieren wir Kernfragen, die wir identifiziert haben, um zwischen den vorhandenen Fairnessdefinitionen zu unterscheiden. Wir beschreiben jeden Punkt im Einzelnen und ergänzen die Ausführungen mit praktischen Beispielen.

7.2.1 Richtlinien

Zielvorgaben für Fairness können über bloße Gleichbehandlung unterschiedlicher Gruppen oder ähnlicher Individuen hinausgehen. Wenn angestrebt wird, benachteiligte Gruppen direkt zu fördern, um bestehende Ungerechtigkeiten auszugleichen, können Maßnahmen wie „positive Diskriminierung“ oder Quoten adäquate Mittel sein. Ein derartiges Ziel kann sich von Gesetzen, Vorgaben von Regulierungsbehörden oder internen Richtlinien einer Organisation ableiten. Dieser Ansatz schließt jeglichen kausalen Zusammenhang zwischen dem sensiblen Attribut und dem Vorhersagewert aus. Falls die vorliegenden Daten in Form von unterschiedlichen Basisraten ein anderes Bild zeichnen, ist das ein starkes Bekenntnis, welches die mathematische Genauigkeit des Algorithmus dieser Überzeugung unterordnet. Wird dieser Entscheidungspunkt bejaht, reduzieren sich die verbleibenden Optionen auf jene Fairnessdefinitionen, die das Prinzip statistischer Unabhängigkeit erfüllen ([Unterabschnitt 5.1](#)).

Zum Beispiel haben es sich viele Universitäten zum Ziel gesetzt, ihre Diversität zu erhöhen, indem sie Studenten und Studentinnen aus benachteiligten Verhältnissen bevorzugt Studienplätze anbieten. Eine derartige Zulassungspolitik gesteht diesen Menschen das gleiche aka-

demische Potential wie Studierenden aus privilegierten Kreisen zu, und macht für ihren eventuell niedrigeren Bildungsstand vielmehr gesellschaftliche Missstände als persönliches Versagen verantwortlich.

7.2.2 Art der Repräsentation

Wenn die Entscheidung im vorangegangenen Punkt zur Beachtung von Richtlinien positiv ausgefallen ist, wird ein besonderer Schwerpunkt auf die ausgewogene Vertretungen der sensiblen Untergruppen gelegt. In diesem Fall gibt es zwei verschiedene Arten von Repräsentation: gleiche Anzahl, unabhängig von den Größen der Untergruppen; oder proportionale Vertretung.

Beispielsweise angenommen, auf eine Stellenausschreibung bewerben sich zehn Frauen und zwei Männer. In Bezug auf das Geschlecht wäre eine Vertretung in gleicher Anzahl gegeben, wenn zwei der Frauen sowie die beiden Männer zum Vorstellungsgespräch eingeladen würden. Um eine proportionale Vertretung zu gewährleisten, müssten fünf Frauen und ein Mann eingeladen werden.

7.2.3 Basisraten

Vorausgesetzt es sind keine Richtlinien zu befolgen, die eine repräsentative Verteilung vorschreiben, dann betrifft die nächste Frage die Basisrate. Diese statistische Kennzahl wurde bereits eingeführt und beschreibt den Anteil der tatsächlich positiven oder negativen Fälle vom gesamten Datensatz (wiederholt in [Abbildung 3](#)). Zwischen den Untergruppen kann die Basisrate gleich ausfallen, oder sie kann abweichen. Wenn die Raten abweichen muss entschieden werden, ob die Fairnessdefinition diesen Unterschied reflektieren

Basisrate


$$BR = \frac{P}{P + N}$$


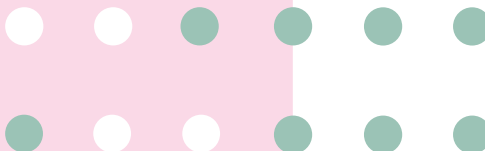
Abbildung 3: Formel und grafische Darstellung der (positiven) Basisrate

soll, oder nicht. Der erste Fall ist gegeben, wenn es einen berechtigten Grund zur Annahme gibt, dass ein Kausalzusammenhang zwischen der Gruppenzugehörigkeit und der Basisrate besteht, und die Fairnessdefinition diesen Effekt berücksichtigen soll. Der zweite Fall gilt, wenn es keine rationale Erklärung gibt, warum die beiden Gruppen grundsätzlich unterschiedliche Ergebnisse liefern sollten. Dann wird die Ursache vielmehr beim Datenerhebungsverfahren vermutet, oder auf andere datenbezogene Aspekte zurückgeführt. Ein weiterer Anlass, von gleichen Basisraten auszugehen, obwohl die Daten andere Rückschlüsse zulassen, liegt vor, wenn für die Abweichung soziale Diskriminierung in der Vergangenheit verantwortlich gemacht wird. Soll die gewünschte Fairnessdefinition eine derartige historische Ungerechtigkeit ausgleichen helfen, stärkt es die Position der unterprivilegierten Gruppe, wenn trotzdem von gleichen Basisraten ausgegangen wird.

In [18] wird diese Frage als zwei gegensätzliche Weltanschauungen definiert: Die Weltanschauung *What you see is what you get* (WYSI-WYG) nimmt an, dass keine strukturellen Verzerrungen in den Daten existieren. Entsprechend geht diese Theorie davon aus, dass statistische Abweichungen in den Basisraten für das Prognosemodell von Relevanz sind und berücksichtigt werden müssen. Dem gegenüber steht die Weltanschauung *We're all equal* (WAE), welche grundsätzlich gleiche Basisraten für alle Gruppen vermutet. Etwaige Abweichungen werden als unerwünschte, strukturelle Verzerrungen interpretiert, die es zu korrigieren gilt.

Wenn davon ausgegangen wird, dass die Basisraten aller Untergruppen identisch sind, kommen als Fairnessdefinitionen nur jene in Frage, die das Prinzip der unbedingten statistischen Unabhängigkeit erfüllen ([Unterabschnitt 5.1](#)). Anderenfalls, wenn die zu implementierende Fairness die unterschiedlichen Basisraten widerspiegeln soll, kommen Fairnessdefinitionen, die Suffizienz oder Separierung erfüllen ([Unterabschnitte 5.2 und 5.3](#)) in Betracht.

In einem medizinischen Kontext zum Beispiel kann festgestellt werden, dass die Basisraten für Männer und Frauen, die an Diabetes lei-



den, etwa gleich sind, wohingegen Brustkrebs in 99% der Fällen bei Frauen auftritt. Eine faire Diagnoseanwendung sollte diesen Unterschied berücksichtigen. Bei Zulassungstests im Rahmen einer Studienplatzvergabe hingegen könnten unterschiedliche Basisraten bei verschiedenen ethnischen Gruppen auf Chancenungleichheiten zurückzuführen sein. Besteht der Anspruch, soziale Ungerechtigkeit auszugleichen, kann hier der passende Ansatz sein, trotzdem gleiche Basisraten anzusetzen.

7.2.4 Ground Truth

ML-Algorithmen lernen von Beispielen. Dabei wird vorausgesetzt, dass die Label der Trainingsdaten den wahren Ausgabewert repräsentieren – sie stellen die sogenannte Ground Truth dar. Da diese Label auch zur Bewertung der Genauigkeit des Prognosemodells herangezogen werden, und sogar eine Referenz für das Fairnessmaß abgeben, falls die gewählte Fairnessdefinition das Label bedingt, spielt die Präsenz einer verlässlichen Ground Truth eine erhebliche Rolle.

Abhängig vom Anwendungsfall kann es allerdings vorkommen, dass keine Ground Truth vorhanden ist, oder dass diese zwar existiert, aber nicht direkt zugänglich ist. Wenn sich das wahre Ergebnis beobachten lässt, besteht eine Ground Truth, und wenn die Label zum Beispiel objektive Messungen oder eindeutige Fakten darstellen, ist der Zugriff meistens nicht weiter schwierig. In anderen Fällen kann das korrekte Ergebnis zwar nicht maschinell erfasst werden, aber es kann von Menschen dennoch eindeutig beobachtet werden, die per sorgfältiger, manueller Zuordnung ausreichend verlässliche Label erzeugen können. Manchmal jedoch ist auch keine Ground Truth vorhanden. Unter diesen Umständen werden die Label abgeleitet und basieren auf menschlichen Entscheidungen, die auf subjektiver Erfahrung beruhen. Solche Label sind möglicherweise vorurteilbehaftet.

Ist keine Ground Truth vorhanden, und lässt sich diese auch nicht auf verlässliche Weise erschließen, dann reduzieren sich die Auswahlmöglichkeiten. Fairnessdefinitionen, die das Prinzip der Separierung ([Unterabschnitt 5.3](#)) erfüllen und die wahre Klasse bedingen, sind in diesem Fall nicht zu empfehlen. Stattdessen kommt eher eine Fairnessdefinition in Frage, die das Prinzip der Unabhängigkeit erfüllt ([Unterabschnitt 5.1](#)) und keine Trainingslabel voraussetzt.

In einem medizinischen Szenario kann für einen Tumor mithilfe einer Biopsie und anschließender Laboruntersuchung abschließend geklärt werden, ob er gut oder bösartig ist. Die Ground Truth ist also verfügbar. In einer Bilderkennungssoftware wiederum, die Tierarten spezifizieren soll, können menschliche Experten Fotos manuell annotieren und so Trainingsdaten von guter Qualität erzeugen. Soll allerdings die Rückfallkriminalität vorhergesagt werden, ist keine Ground Truth unmittelbar verfügbar, da mögliche neue Straftaten erst in der Zukunft stattfinden, und außerdem gar nicht jedes Verbrechen aktenkundig wird.

7.2.5 Erklärende Variablen

Unter Umständen gibt es Attribute, die als legitime Quelle für Unterschiede in den Daten betrachtet werden können. Wenn aufgezeigt werden kann, dass eine Verzerrung in den Daten zwischen den Untergruppen auf diese Variablen zurückzuführen ist, kann dieser Unterscheid begründet und die Abweichung akzeptiert werden [7].

Angenommen es sollen Gehaltsklassen für Jobbewerber geschätzt werden. Im vorliegenden Datensatz arbeitet eine Gruppe aber im Durchschnitt weniger Arbeitsstunden als die andere. Dann könnte ein Attribut *working_hours* als erklärende Variable dienen.

7.2.6 Verzerrte Label

Wenn die Ground Truth nicht direkt zugänglich ist, und die vorhandenen Label von Menschen manuell zugeordnet wurden, besteht das Risiko, dass die so erhobenen Daten gewisse Verzerrungen enthalten. Die Label werden zur Bestimmung der Genauigkeit des Prognosemodells herangezogen, und spielen auch für die Bewertung der Fairness eine Rolle, falls deren Definition die Label berücksichtigt. Daher ist es entscheidend, möglichen Quellen von Verzerrung entgegenzuwirken, zum Beispiel mit einem Labelkorrektur-Framework [19, 20]. Sollte dieser Vorgang keine zufriedenstellenden Ergebnisse liefern, ist keine belastbare Ground Truth vorhanden, und es gilt die gleiche Argumentation wie zuvor: Es ist nicht ratsam, eine Fairnessdefinition zu verwenden, welche sich auf den wahren Ausgabewerte stützt, sondern besser eine Definition zu wählen, die das Prinzip der Unabhängigkeit erfüllt (Unterabschnitt 5.1).

Wenn zum Beispiel eine Software lernen soll, Fotos mit Worten zu beschreiben, dann erzeugen Menschen die zugehörigen Trainingsdaten, indem sie Beispielbilder verschlagworten. Diese Aufgabe lässt einen gewissen kreativen Freiraum zu, etwa durch die Auswahl der Objekte, oder deren Bezeichnung. Insbesondere wenn diese Tätigkeit nur von einer kleinen Gruppe Menschen ausgeübt wird, können sich in den Trainingsdaten Verzerrungen niederschlagen.



Abbildung 4: Formeln und symbolische Darstellung beider Metriken

7.2.7 Relevanz und Sensitivität

Wenn feststeht, dass eine ausreichend verlässliche Ground Truth vorhanden ist, muss als nächstes über ein bekanntes Problem beim Maschinellen Lernen entschieden werden: der Kompromiss zwischen Relevanz und Sensitivität. Relevanz (*precision*) beschreibt den Anteil der positiven Vorhersagen, der korrekt ist – zuvor eingeführt als positiver Vorhersagewert (PPV). Sensitivität (*recall*) beschreibt den Anteil der tatsächlich positiven Fälle, der korrekt vorhergesagt wurde – in diesem Bericht auch als Richtig-Positiv-Rate (TPR) bezeichnet. In [Abbildung 4](#) stellen wir beide Kennzahlen erneut mathematisch und symbolisch dar. Zu klären ist nun, welches der beiden Maße für den gegebenen Anwendungsfall in Bezug auf die Fairness eine kritischere Rolle spielt. Als Faustregel gilt, dass wenn die Konsequenzen für den oder die Betroffene im Zweifelsfall eine negative, strafende Auswirkung haben, dann der Schwerpunkt bezüglich Fairness auf der Relevanz liegen sollte. Ist das Ergebnis im besten Fall vielmehr vorteilhaft, im Sinne von

Unterstützung, auf die die Person sonst verzichten müsste, dann ist oft der Aspekt der Sensitivität wesentlicher in puncto Fairness. Die Antwort auf diese Frage legt fest, aus welcher der verbleibenden beiden Kategorien die endgültige Fairnessdefinition stammen wird: Wenn der Schwerpunkt auf gleichen Relevanzraten für beide Untergruppen liegt, dann wird die Fairnessdefinition auf der vorhergesagten Klasse beruhen, und folglich dem Prinzip der Suffizienz genügen ([Unterabschnitt 5.2](#)). Anderenfalls, wenn der Fokus auf der Sensitivität liegt, wird die resultierende Fairnessdefinition auf der wahren Klasse basieren und das Separierungs-Prinzip erfüllen ([Unterabschnitt 5.3](#)).

In einem Anwendungsfall, bei dem es um Betrugserkennung bei Versicherungsfällen geht, könnte man es als oberstes Fairnessziel betrachten, die Zahl der zu Unrecht als betrügerisch eingeordneten Fälle zu minimieren und für beide Untergruppen die Relevanzraten gleich niedrig zu halten. In einem Kreditvergabeszenario hingegen könnte der Fokus in Bezug auf Fairness bei der Sensitivität liegen. Das hieße, kreditwürdigen Bewerbern und Bewerberinnen aus beiden Untergruppen sollten ihre Anträge zu gleich hohen Raten bewilligt werden.

7.2.8 Ausgabearten

Eine mehr praktische als ethische Frage, aber dennoch relevant um die finale Fairnessdefinition zu wählen, ist jene nach der gewünschten Ausgabeart. Ein Score ist ein kontinuierlicher Wert. Oftmals liegt er zwischen 0 und 1, und beschreibt dann die Wahrscheinlichkeit, dass

Falsch-positiv-Rate

$$FPR = \frac{FP}{N} = \frac{\text{[red semi-circle]}}{\text{[red + grey semi-circles]}}$$

Falsch-negativ-Rate

$$FNR = \frac{FN}{P} = \frac{\text{[green semi-circle]}}{\text{[green + black semi-circles]}}$$

Abbildung 5: Formeln und symbolische Darstellungen der Fehlerarten

der gegebene Fall der positiven Klasse angehört. Ist die Ausgabeart stattdessen ein Label, entspricht das Ergebnis eindeutig einer der beiden Klassen.

In einem Kreditvergabeszenario wird oft ein Score bevorzugt, weil sein Wert mehr Interpretationsspielraum für den Mensch lässt, der die finale Entscheidung trifft. Wird das Ergebnis allerdings automatisch weiterverarbeitet, zum Beispiel in einem Online-Marketing Szenario, könnten Klassenlabel als Ausgabeart sinnvoller sein.

7.2.9 Fehlerarten

Die letzte Entscheidung betrifft schließlich die Frage, welche Fehlerart im gegebenen Anwendungsfall eine höhere Priorität bezüglich Fairness hat. Die unterschiedlichen zu berücksichtigenden Fehlerarten sind die Falsch-positiv und die Falsch-negativ-Rate (wie bereits etwas früher eingeführt und in [Abbildung 5](#) rekapituliert). Beide Kennzahlen bemessen Fehlklassifikation, abhängig vom Einsatzgebiet kann eine Fehlerart allerdings eine bedeutsamere Rolle beim Erreichen von Fairness haben, als die andere. Generell gilt, dass für Hochrisiko-Anwendungen sowohl die Falsch-positiv als auch die Falsch-negativ-Rate für alle Gruppen auf gleichem Niveau gehalten werden sollte. Für weniger sicherheitskritische Anwendungen könnte das Fairnessziel zugunsten erhöhter technischer Flexibilität etwas abgeschwächt werden, indem ein überschaubares zusätzliches Risiko in Kauf genommen wird [\[12\]](#). Um in dieser Angelegenheit einen besseren Überblick zu haben, kann es nützlich sein, die Wahrheitsmatrix (siehe [Unterabschnitt 3.2](#)) um eine Beschreibung der Ereignisse im Fehlerfall zu ergänzen. So lassen sich die Konsequenzen einer korrekten oder inkorrekten Einordnung vor Augen führen und entsprechend gewichten.

In einem Online-Marketing Szenario, wo eine Stellenanzeige Männer und Frauen mit relevanten Profilen eingeblendet werden soll, mögen Unterschiede in der Falsch-positiv-Rate (die Anzeige also Menschen zu zeigen, die eigentlich nicht für die Stelle in Frage kommen) zwischen den Untergruppen verkraftbar sein, solange die Anteile der Leute mit relevanten Profilen gleichermaßen hoch sind. Bei einer Gesichtserkennungssoftware andererseits sollten beide Fehlerarten für sämtliche Hauttypen gleich niedrig sein.

7.3 Beispielanwendung

Wir testen den *Fairness Compass* für unser wiederkehrendes Beispiel zur Betrugserkennung bei Versicherungsfällen. Mit unserem interaktiven Tool lässt sich die Entscheidungsfindung für die Wahl einer Fairnessdefinition einfach und transparent darstellen (siehe [Online-Tool²](#)). Wohlgemerkt handelt es sich bei diesem Beispiel um ein Gedankenexperiment. Für dasselbe Szenario sind verschiedene Argumente denkbar, die zu einem anderen Ergebnis führen. Der Zweck des *Fairness Compass* ist nicht, eine Lösung vorzuschreiben, sondern vielmehr die Entscheidungsfindung zu strukturieren und das Ergebnis argumentativ zu rechtfertigen.

7.4 Weitere Entwicklung

Die Forschung im Bereich KI und Fairness schreitet stetig voran. Wahrscheinlich werden neue Fairnessdefinitionen entwickelt werden. Die allgemeine Fairness-Debatte wird sicherlich fortgesetzt werden, und die Gesellschaft wird konkretere Erwartungen formulieren, wie faire Entscheidungen in bestimmten Anwendungsbereichen auszuweisen haben. Um diese zukünftigen Entwicklungen zu berücksichtigen, haben wir bei unserer technischen Architektur den Schwerpunkt auf Erweiterbarkeit und Anpassbarkeit gelegt. Das Online-Tool wurde mit der kostenlosen Online-Diagrammsoftware [diagrams.net](#) realisiert. Wir haben damit den Entscheidungsbaum entworfen und online veröffentlicht. Das Schema ist im XML-Format gespeichert, was eine Versionierung und Nachverfolgung von Änderungen und Erweiterungen möglich macht. Wir haben außerdem ein Plug-In für [diagrams.net](#) implementiert, das dessen Funktionsumfang um die oben beschriebenen interaktiven Features erweitert. Den [Quellcode³](#) für dieses Plugin haben wir ebenfalls online zur Verfügung gestellt.

2 <https://axa-rev-research.github.io/fairness-fraud-study.html>

3 <https://axa-rev-research.github.io/drawio/src/main/webapp/plugins/props.js>



8. Schlussfolgerungen

Dieser Bericht behandelt das Problem von Verzerrungen in Daten und KI-Anwendungen. Wir erklären seine Ursachen und Konsequenzen, und führen zudem aus, warum es für die Lösung keinen Königsweg gibt. Zur allgemeinen Übersicht stellen wir die verfügbaren Fairnessdefinitionen für Klassifizierungsprobleme vor, und erläutern deren unterschiedlichen Eigenschaften anhand von Beispielen. Als Orientierungshilfe und praktischen Lösungsansatz präsentieren wir den *Fairness Compass* – ein Tool in Form eines Entscheidungsbaums, das einige wesentliche Fragen zur gewünschten Art von Fairness abfragt, und dann auf Grundlage der Antworten für einen konkreten Anwendungsfall die am besten passende Option liefert. Dieses Tool ist auch nützlich, um die Argumente bei der Entscheidungsfindung zu dokumentieren. Menschen, die von der Entscheidung eines KI-Systems betroffen sind, können so einfacher nachvollziehen, warum die vorliegende Art von Fairness implementiert wurde. Erhöhte Transparenz in dieser Form kann dazu beitragen, das Vertrauen in KI-Anwendungen zu stärken.

Wir möchten betonen, dass das hier vorgestellte Schema wohl nicht das letzte Wort zu diesem Thema sein wird. Die Forschung wird neue Erkenntnisse liefern, und die gesellschaftlichen Erwartungen werden sich konkretisieren. Wir verstehen unsere Arbeit daher als ersten Schritt, das komplexe aktuelle Angebot von Fairnessdefinitionen zu strukturieren. Wir würden uns freuen, wenn dieses Projekt zu fundierten Entscheidungen in konkreten Anwendungsszenarien beiträgt, und hoffen weiterhin, dass es als Basis für grundsätzliche Diskussionen dient, und so einen nützlichen Beitrag für die Implementierung von mehr Fairness in realen KI-Systemen leistet.

Danksagungen

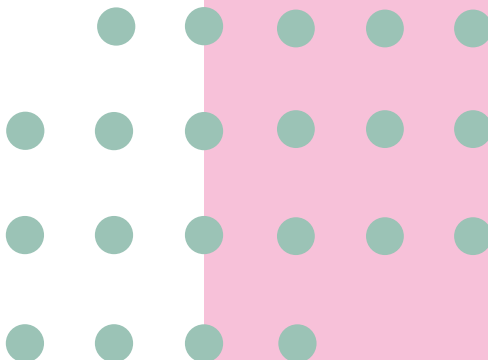
Wir bedanken uns bei Jonathan Aigrain für die anregenden Diskussionen und das konstruktive Feedback zu diesem Dokument.

9. Literatur

- [1] Sam Corbett-Davies and Sharad Goel. The measure and mis-measure of fairness: A critical review of fair machine learning. CoRR, abs/1808.00023, 2018.
- [2] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. CoRR, abs/1908.09635, 2019.
- [3] Council of Europe. Charter of fundamental rights of the european union. (2012/C 326/02).
- [4] Solon Barocas and Andrew D Selbst. Big data’s disparate impact. Calif. L. Rev.. California Law Review, 104(IR):671.
- [5] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. Fairness Through Awareness. 2011.
- [6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and Machine Learning. fairmlbook.org, 2019. <http://www.fairmlbook.org>.
- [7] F. Kamiran, I. Zliobaite, and T.G.K. Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. Knowledge and Information Systems, 35(3):613–644, 2013.

- [8] Pedro Saleiro, Benedict Kuester, Abby Stevens, Ari Anisfeld, Loren Hinkson, Jesse London, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. CoRR, abs/1811.05577, 2018.
- [9] Richard Berk. A primer on fairness in criminal justice risk assessments. *The Criminologist*, 41(6):6–9, 2016.
- [10] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5, 10 2016.
- [11] Cynthia Crowson, Elizabeth J. Atkinson, Terry M Therneau, Andrew B. Lawson, Duncan Lee, and Ying MacNab. Assessing calibration of prognostic risk scores. *Statistical Methods in Medical Research*, 25(4):1692–1706, August 2016.
- [12] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of Opportunity in Supervised Learning. pages 1–22, 2016.
- [13] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17*, page 797–806, New York, NY, USA, 2017. Association for Computing Machinery.
- [14] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. CoRR, abs/1609.05807, 2016.
- [15] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 03 2017.
- [16] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, May 23, 2016, 2016.

- [17] William Dieterich, Christina Mendoza, and Tim Brennan. Com-pas risk scales: Demonstrating accuracy equity and predictive parity. Northpointe Inc, 2016.
- [18] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkata-subramanian. On the (im)possibility of fairness. Sep 2016.
- [19] Michael Wick, Swetasudha Panda, and Jean-Baptiste Tristan. Unlocking fairness: a trade-off revisited. In H. Wallach, H. La-rochelle, A. Beygelzimer, F. d’Alché Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems 32, pages 8780–8789. Curran Associates, Inc., 2019.
- [20] Heinrich Jiang and Ofir Nachum. Identifying and correcting la-bel bias in machine learning. CoRR, abs/1901.04966, Jan 2019.



Expert



Boris RUF boris.ruf@axa.com

Research Data Scientist

Sponsor



Marcin DETYNIECKI marcin.detyniecki@axa.com

Group Chief Data Scientist and Head of AI Research & Thought Leadership

Welche Art von Fairness macht KI-Systeme gerecht?

Juni 2021

Herausgegeben von GETD | AI Research & Thought Leadership

61 rue Mstislav Rostropovitch
75017 Paris, France
marcin.detyniecki@axa.com

Design & Artwork

Viviane Badach

Für dieses Dokument wurden die Schriftarten Source Sans Pro und Publico verwendet.

© 2021 AXA. Alle Rechte vorbehalten.

Diese Broschüre ist unter dem Titel „Towards the Right Kind of Fairness in AI“ auch auf Englisch verfügbar.



