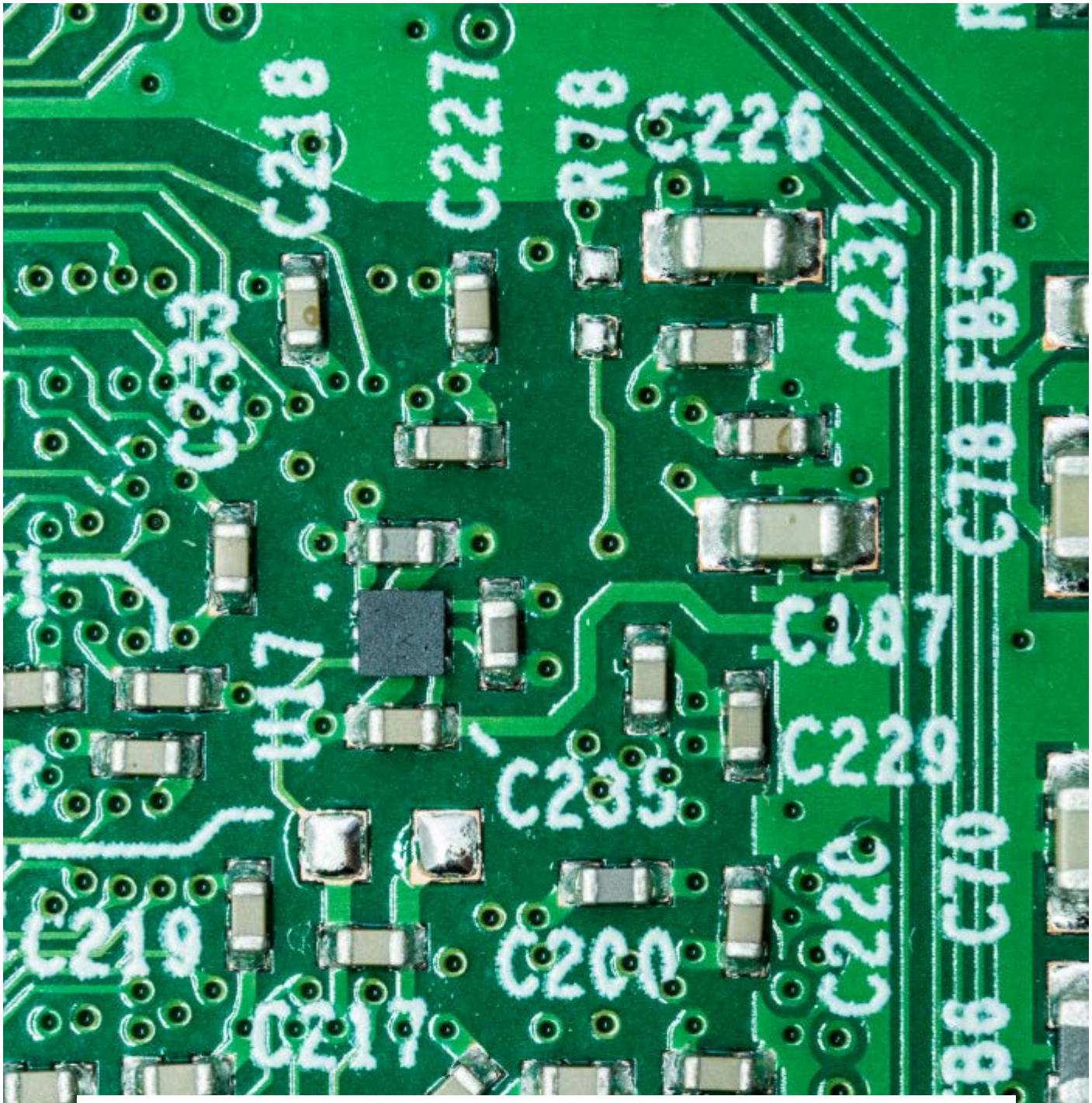

Wenn Maschinen Menschen bewerten

Internationale Fallbeispiele für Prozesse algorithmischer Entscheidungsfindung

- Arbeitspapier -

BertelsmannStiftung

Wenn Maschinen Menschen bewerten

Internationale Fallbeispiele für Prozesse algorithmischer Entscheidungsfindung
- Arbeitspapier -

Impressum

Bertelsmann Stiftung
Carl-Bertelsmann-Straße 256
33311 Gütersloh
www.bertelsmann-stiftung.de

Verantwortlich
Konrad Lischka
Ralph Müller-Eiselt

Autoren
Konrad Lischka
Anita Klingel

DOI 10.11586/2017025

Inhalt

1 Vorwort

Bis zu 70 Prozent der Stellenbewerber in Großbritannien und den Vereinigten Staaten werden zuerst von automatisierten algorithmischen Verfahren bewertet, bevor ein Mensch ihre Unterlagen sieht (Weber und Dwoskin 2014). Gerichte in neun US-Bundesstaaten nutzen in Strafverfahren Software, die Risikoprognosen für die Angeklagten berechnet (Angwin et al. 2016: 2). Automatisierte Prognosen zur Kreditwürdigkeit werden in den Vereinigten Staaten auch genutzt, um die Höhe von Versicherungspolicen zu bestimmen. Das FBI gleicht Bildmaterial von Straftätern automatisiert mit 411 Millionen Bildern aus Führerschein-, Pass- und Visadaten ab, um mögliche Verdächtige zu identifizieren.

Diese vier Beispiele zeigen: Menschen werden heute in vielen Lebensbereichen von Prozessen algorithmischer Entscheidungsfindung bewertet, sogenanntem „algorithmic decision making“ (ADM) (Zweig 2016). Solche ADM-Prozesse sind seit Jahren im Einsatz und kategorisieren Menschen ohne große Debatte über Fairness, Erklärbarkeit, Überprüfbarkeit oder Korrigierbarkeit der Verfahren. Das kann daran liegen, dass die Systeme wenig mit den künstlichen Intelligenzen (KI) aus der Science-Fiction verbindet. Menschen assoziieren mit KI oft Eigenschaften fiktionaler Figuren wie HAL 9000 oder Wintermute: Intentionalität und Bewusstsein. Solche starken KI existieren aber bislang nur in Literatur und Film. Nichts davon zeichnet die Systeme aus, die wir in dieser Fallsammlung vorstellen. Und doch haben sie bereits erheblichen Einfluss vor Gericht, bei der Vergabe von Krediten und Studienplätzen, dem Einsatz von Polizeikräften, der Berechnung von Versicherungstarifen und der Aufmerksamkeit, die Anrufer in einer Kundendienst-Hotline erfahren. Es sind allesamt auf bestimmte Probleme spezialisierte Programme, die das Leben vieler Menschen beeinflussen. Es geht nicht um Science-Fiction, sondern um die Gegenwart (Lischka 2015).

Die in diesem Arbeitspapier aufbereiteten Fallbeispiele zeigen Chancen und Risiken solcher Prozesse. Chancen wie diese: Mustererkennung kann dabei helfen, das Risiko von Bleivergiftungen bei Kindern abhängig vom Wohnort vorherzusagen (siehe Kapitel 2.3 Prognose drohender Bleivergiftungen) oder Hotspots für bestimmte Delikte zu prognostizieren (zum Beispiel Wohnungseinbrüche, siehe Kapitel 2.8 ortsbezogene Kriminalitätsprognosen). Ein künstliches neuronales Netz rechnet anhand von Satellitenfotos die regionale Verteilung von Armut in Entwicklungsländern fast genauso gut hoch wie erheblich teurere Umfragen vor Ort. Diese Ergebnisse könnten dazu genutzt werden, Armut zielgerichtet dort zu bekämpfen, wo Not und folglich die Wirkung von Hilfsmaßnahmen am größten sind (siehe Kapitel 2.4 Armutsverteilung vorhersagen).

Um diese Chancen für mehr Teilhabe zu nutzen, müssen Prozesse algorithmischer Entscheidungsfindung bei der Planung, Gestaltung und Umsetzung klar auf dieses Ziel ausgerichtet werden. Wenn dies nicht geschieht, kann der Einsatz dieser Werkzeuge aber auch ohne Weiteres zu mehr sozialer Ungleichheit führen. Die in den ausgewählten Fallbeispielen erkennbaren Risiken und Fehlentwicklungen zeigen Fehlerquellen auf, die bei vielen ADM-Prozessen auftreten können. Nicht selten sind bei einzelnen Anwendungsszenarien mehrere dieser Mängel zu beobachten. Wir heben in diesem Arbeitspapier dennoch an jedem Fallbeispiel einen typischen Handlungsbedarf besonders hervor, der bei der künftigen Gestaltung von ADM-Prozessen für mehr Teilhabe beachtet werden sollte.

Tabelle 1: Handlungsbedarf bei Prozessen algorithmischer Entscheidungsfindung (Quelle: eigene Darstellung)

Handlungsbedarf	Beschreibung	Fallbeispiel
Falsifizierbarkeit sicherstellen	ADM-Prozesse können asymmetrisch aus Fehlern lernen. Asymmetrisch bedeutet: Das System kann qua Design des gesamten Prozesses nur bestimmte Arten der eigenen Fehlprognosen nachträglich erkennen. Wenn Algorithmen asymmetrisch lernen, drohen selbstverstärkende Rückkoppelungseffekte.	Rückfallprognosen vor Gericht
Sachgerechte Anwendung sichern	Institutionslogik kann dazu führen, dass ADM-Prozesse völlig anders eingesetzt werden, als es die Entwickler vorgesehen haben. Derart unsachgerechter Einsatz ist zu vermeiden.	Individuelle Kriminalitätsprognosen
Geeignete Wirkungslogik finden	Durch Algorithmen ermöglichte Effizienzgewinne einzelner Prozessschritte können die Frage überdecken, ob die zur Lösung eines gesellschaftlichen Problems eingesetzten Mittel insgesamt angemessen sind.	Prognose drohender Bleivergiftungen
Konzepte korrekt messbar machen	Soziale Phänomene oder Konzepte, wie zum Beispiel Armut oder der soziale Ungleichheit, sind häufig schwer zu operationalisieren. Hilfreich sind im öffentlichen Diskurs entwickelte und belastbare Kennzahlen.	Armutsverteilung vorhersagen
Umfassende Evaluation gewährleisten	Die normative Kraft des technisch Machbaren überholt allzu leicht die Diskussion über das gesellschaftlich Sinnvolle. So kann die Skalierbarkeit maschineller Entscheidungen schnell zu Einsatzszenarien führen, deren gesellschaftliche Angemessenheit und Folgen nicht geprüft und nicht debattiert worden sind.	Automatische Gesichtserkennung
Vielfalt von ADM-Prozessen sichern	Die einmal entwickelte Entscheidungslogik eines ADM-Prozesses ist auf sehr viele Fälle anwendbar, ohne dass die Kosten für den Einsatz substanziell steigen. Das führt dazu, dass in einigen Lebensbereichen wenige ADM-Verfahren dominieren können. Je größer die Reichweite ist, desto schwieriger ist es für den Einzelnen, sich der Verfahren und Folgen zu entziehen.	Bewerbervorauswahl per Online-Persönlichkeitstests
Überprüfbarkeit ermöglichen	Ob ein ADM-Prozess ein adäquates Konzept von Fairness verwendet, wird häufig nicht überprüft. Wenn Logik und Natur eines Algorithmus geheim gehalten werden, ist dies sogar unmöglich. Ohne Überprüfung durch unabhängige Dritte kann keine informierte Debatte über Chancen und Risiken eines spezifischen ADM-Prozesses geführt werden.	Studienplatzvergabe in Frankreich
Soziale Wechselwirkungen beachten	Selbst bei einem sehr eingeschränkten Einsatz ist die Wechselwirkung von ADM-Prozessen mit der Umwelt sehr komplex. Nur die Analyse der Wirkungen des gesamten sozioinformatischen Prozesses kann zeigen, in welchem Verhältnis Chancen und Risiken stehen	Ortsbezogene Kriminalitätsprognosen
Zweckentfremdung verhindern	Leicht abrufbare Prognosen wie Scoringwerte können für nicht angemessene Ziele genutzt werden. Solche Zweckentfremdungen sind unbedingt zu verhindern.	Kreditscoring in den USA

Die hier beschriebenen Fallbeispiele sind ein Mittel, um an konkreten, zum Teil evaluierten ADM-Prozessen Chancen und Risiken herauszuarbeiten und zu abstrahieren. Dieses Dokument spiegelt einen ersten Zwischenstand unserer Auseinandersetzung mit dem Thema wider. Wir veröffentlichen es als Arbeitspapier, um einen Beitrag zu einem sich schnell entwickelnden Feld zu geben, auf dem auch andere aufbauen können. Daher veröffentlichen wir das Arbeitspapier unter einer freien Lizenz (CC BY-SA 3.0 DE), damit es beispielsweise auch als Diskussionsgrundlage für Workshops oder andere Auseinandersetzungen mit der Materie genutzt werden kann.

Algorithmische Entscheidungsfindung wird nur dem Wohl der Gesellschaft dienen, wenn sie diskutiert, kritisiert und korrigiert wird. Es ist Zeit für diesen Diskurs in Deutschland. Wir haben jetzt die Chance, von internationalen Beispielen und Erfahrungen zu lernen und eine Entwicklung zu gestalten, die insbesondere in den Vereinigten Staaten schon deutlich weiter ist: Dort hat das Weiße Haus noch unter Präsident Barack Obama einen Bericht zu den Herausforderungen durch maschinelle Entscheidungen vorgelegt (Executive Office of the President, National Science and Technology Council und Committee on Technology 2016). In Deutschland sind ADM-Prozesse noch nicht so präsent. Deutsche Gerichte nutzen keine ADM-Risikoprognosen. Erst 60 der 1000 größten Unternehmen hierzulande haben 2016 computergesteuerte Verfahren zur Bewerberauswahl verwendet (Eckhardt et al. 2016: 8). Und automatisierte Gesichtserkennung ist in Deutschland nur an sieben Flughäfen in das Grenzkontrollsystem EasyPASS eingebunden (Bundespolizei 2015).

Noch können wir also bestimmen, wie wir als Gesellschaft Algorithmen einsetzen wollen. Wie die Initiative Algorithmwatch es formuliert: „Wir müssen entscheiden, wie viel unserer Freiheit wir an ADM übertragen wollen“ (Algorithmwatch 2016). Dabei sollten wir dabei nicht nur das Wie, sondern an einigen Stellen auch das Ob diskutieren: Wo die Gesellschaft sich zum Beispiel für Solidarität und Vergemeinschaftung von Risiken entschieden hat, dürfen ADM-Prozesse diese Risiken nicht individualisieren. Nicht das technisch Mögliche, sondern das gesellschaftlich Sinnvolle muss Leitbild sein – damit maschinelle Entscheidungen den Menschen dienen.

Ralph Müller-Eiselt
Senior Expert
Taskforce Digitalisierung
Bertelsmann Stiftung

Konrad Lischka
Project Manager
Taskforce Digitalisierung
Bertelsmann Stiftung

2 Fallbeispiele

Es folgen neun Beispiele für den Einsatz von ADM-Verfahren. Die Reise beginnt in den Vereinigten Staaten, mit Anwendungen, die uns in dieser Form nur dort bekannt sind, etwa Prognosen des Rückfallrisikos Angeklagter vor Gericht oder zum Risiko von Bleivergiftungen in Chicago. Nach einem transnationalen Beispiel (Auswertung von Satellitenfotos für Kartierung von Armut) nähern wir uns anschließend Europa: Ein Fallbeispiel aus Frankreich (Hochschulzugang) und einige auch in Deutschland angewendete US-Verfahren (z. B. ortsbezogenes „predictive policing“) zeigen, dass der Einsatz von ADM-Verfahren ein weltweites Phänomen ist, das auch hierzulande um sich greift. Jede Fallbeschreibung hebt eine typische Fehlerquelle und damit einen Handlungsbedarf bei der künftigen Gestaltung von ADM-Prozessen für mehr Teilhabe besonders hervor. Diese Pointierung erfolgt so trennscharf wie möglich, aber im vollen Bewusstsein, dass die identifizierten Mängel in der Praxis nicht selten leider gehäuft auftreten.

Die Darstellung der einzelnen Fallbeispiele haben wir einheitlich strukturiert, um mit einem schnellen Überblick über Fakten und Einschätzungen eine Basis für Diskussionen zu liefern. Einer knappen Zusammenfassung folgt jeweils die Beschreibung des Outputs der Systeme sowie der ihm zugrunde liegenden Datenbasis und Entscheidungslogik. Anschließend stellen wir die Konsequenzen und vorliegenden Evaluationsergebnisse der Verfahren dar, um nicht nur die Technik, sondern den gesamten sozioinformatischen Prozess zu betrachten. Der Abschnitt „Zur Diskussion“ fasst jeweils knapp die in den Fachdebatten thematisierten Chancen und Risiken zusammen. Wenn zu den geschilderten Fallbeispielen Parallelen in Deutschland bekannt sind, skizzierten wir diese kurz im abschließenden Abschnitt „Situation und Relevanz in Deutschland“.

2.1 Falsifizierbarkeit sicherstellen: Rückfallprognosen vor Gericht

Software berechnet Prognosen der Rückfallwahrscheinlichkeit von Straftätern. Solche algorithmisch unterstützten Entscheidungsprozesse sind heute in fast jedem US-Bundesstaat an zumindest einem Punkt in strafrechtlichen Verfahren im Einsatz (Barry-Jester, Casselman und Goldstein 2015). Mehr als 60 Prognoseinstrumente sind auf dem Markt, viele kommen von Unternehmen, darunter das weit verbreitete System COMPAS der Firma Northpointe.

2.1.1 Output: Risikoprognosen und Hilfsbedarfe auf einer Skala von 1 bis 10

Das COMPAS-System bewertet Begutachtete in unterschiedlichen Kategorien, die in insgesamt 43 Skalenwerten bemessen werden (Northpointe 2015: 2). Das sind zum einen Risikoprognosen: zum Beispiel das allgemeine Rückfallrisiko oder das Rückfallrisiko bei spezifischen Gewalttaten. Andere Kategorien sollen Hilfsbedarfe der Bewerteten erkennen und quantifizieren, um die Planung von Interventionen zu systematisieren. Dazu dienen Kategorien wie Armut, Süchte, kriminelle Verwandtschaft. Alle Ausprägungen werden auf einer Skala von 1 bis 10 ausgegeben. Bei einer Risikoprognose von 1 bis 4 sei die Rückfallwahrscheinlichkeit „niedrig“, bei 5 bis 7 „mittel“ und bei 8 bis 10 „hoch“, heißt es im Anwenderhandbuch (a.a.O.: 11).

Wenn Richtern diese Prognosen vorliegen, entscheiden sie, ob und wie sie sie in ihr Urteil einbeziehen.

2.1.2 Datenbasis und Entscheidungslogik: Vergleich mit Testergebnissen einer Normgruppe

Die Scoringwerte leitet das COMPAS-Verfahren aus Antworten auf 137 Fragen ab. Die Angaben kommen aus Polizeiakten und aus von den Bewerteten ausgefüllten Fragebögen. Enthalten sind Fragen wie: War einer Ihrer Elternteile je in Haft? Wie viele Ihrer Freunde/Bekannten konsumieren illegal Drogen? Die Begutachteten sollen zudem Aussagen bewerten wie diese: „Eine hungrige Person hat das Recht zu stehlen“ (Angwin et al. 2016: 3).

Welche Antworten bei der Berechnung wie gewichtet werden, ist nicht öffentlich bekannt. Im Anwenderhandbuch skizziert das Unternehmen die Faktoren nur sehr vage. So heißt es zur Berechnung der Prognose des allgemeinen Rückfallrisikos: „The primary factors making up this scale involve prior criminal history, criminal associates, drug involvement, and early indicators of juvenile delinquency problems“ (Northpointe 2015: 30).

Der für Richter und Vollzugsbeamte sichtbare COMPAS-Output auf einer Skala von 1 bis 10 gibt einen Vergleich zur Verteilung der Werte in einer Normgruppe an. Diese Normgruppe besteht aus 7381 Straftätern, die 2004 und 2005 in US-Gefängnissen mit dem COMPAS-Verfahren untersucht worden sind (a.a.O.: 14). Die eigentlich ermittelten Scoringwerte in den 21 Kategorien wurden nach Dezilen aufgeteilt. So erzielte zum Beispiel in der Normgruppe ein Zehntel der Bewerteten einen Wert von 23 oder weniger in der Kategorie „kriminelle Persönlichkeit“, das zweite Dezil erzielte Werte von 24 bis 25 und so weiter. Die COMPAS-Werte geben also an, welchem Dezil der Normgruppe ein Bewerteter aufgrund seiner Scoringwerte ähnelt. Die einzelnen Skalen unterscheiden sich, die Dezilangabe macht die Bewertung handhabbar.

2.1.3 Konsequenzen: Je höher die Risikoprognose, desto wahrscheinlicher ist Haft

Gerichte in vielen US-Bundesstaaten greifen bei Entscheidungen über Kautionen oder vorzeitige Haftentlassung auf solche Verfahren zurück. Pennsylvania entwickelt ein Verfahren, um auch bei Verurteilungen Risikoprognosen einzubeziehen (Pennsylvania Commission on Sentencing 2016). In neun US-Bundesstaaten liegen solche Prognosen Richtern bei Strafverfahren vor (Angwin et al. 2016: 2). In einigen Fällen begründeten Richter Urteile mit COMPAS-Prognosen. Im Februar 2013 wurde Eric Loomis in Wisconsin verhaftet, weil er ein Auto fuhr, das zuvor bei einer Schießerei genutzt wurde. Loomis bekannte sich schuldig, sich den Polizeibeamten entzogen zu haben. Er wurde zu achteinhalb Jahren Haft verurteilt. Der Richter erklärte, der Angeklagte sei durch die COMPAS-Bewertung als ein großes Risiko für die Gemeinschaft identifiziert worden (a.a.O.: 10).

2.1.4 Evaluation: Unterschiedliche Fehlprognosen für Schwarze und Weiße

Den Einsatz von COMPAS im County Broward in Florida hat das durch Stiftungen finanzierte US-Rechercheorganisation Propublica 2016 untersucht. Die Reporter haben Prognosen des Rückfallrisikos von 7000 in den Jahren 2013 und 2014 Verhafteten ausgewertet. Sie prüften, ob die Personen in den folgenden zwei Jahren wegen neuer Straftaten angeklagt worden sind. Kernergebnisse der Propublica-Recherche:

- 20 Prozent der Personen mit einer Rückfallprognose für Gewaltkriminalität wurden binnen zwei Jahren nach der Prognose wegen eines solchen Delikts angeklagt (Angwin et al. 2016: 2).
- 61 Prozent der Personen (etwas besser als der Zufall) mit einer allgemeinen Rückfallprognose wurden in den zwei Folgejahren wieder aktenkundig – Ordnungswidrigkeiten eingerechnet (ebd.).
- Die Art der Fehlprognosen unterscheidet sich zwischen schwarzen und weißen Personen: Der Anteil Schwarzer mit hoher Rückfallprognose aber ohne Rückfall binnen zwei Jahren ist doppelt so hoch wie der Weißer (ebd.).

Diese Ergebnisse verteidigen einige Autoren als fair, da die Rückfallquote insgesamt bei schwarzen Angeklagten signifikant höher sei als bei weißen: „Racial differences in failure rates across race describe the behavior of defendants and the criminal justice system, not assessment bias“ (Flores, Bechtel und Lowenkamp 2016: 13)

Für andere in den Vereinigten Staaten genutzte Verfahren zur Rückfallprognose gibt es kaum unabhängige Evaluationen. Die Validität wurde in den meisten Fällen lediglich in „ein oder zwei Studien untersucht und häufig stammen diese Studien von denselben Leuten, die das Instrument entwickelt haben“ (Desmarais und Singh 2013: 53). Hinzu kommt: Fast alle Studien untersuchen lediglich, ob die Verfahren bekannten Rückfalltätern ein höheres Rückfallrisiko prognostizieren, nicht aber, ob Personen mit hoher Rückfallprognose tatsächlich rückfällig werden (a.a.O.: 55).

2.1.5 Zur Diskussion: Ohne Falsifikation droht Verzerrung

Das Fallbeispiel veranschaulicht ein Kernproblem, das bei vielen Risikoprognosen auftreten kann: Sie können in einer Feedbackschleife die eigene Entscheidungsgrundlage verzerren. Bei Rückfallprognosen vor Gericht zum Beispiel so: Nehmen wir an, Richter tendieren bei höheren Risikoprognosen dazu, eher Haft als Bewährung anzuordnen. Durch den längeren Gefängnisaufenthalt kann die Wahrscheinlichkeit eines Rückfalls steigen, weil Menschen zum Beispiel in neue, kriminelle soziale Kontexte integriert werden. Das kann dazu führen, dass bei Menschen mit hoher Risikoprognose die Rückfallquote nach der Haft tatsächlich höher ist. So könnte sich das Prognosesystem selbst bestätigen (die Risikoprognose trifft zu) und unter Umständen langfristig sogar selbst bestärken, wenn das System auf neuen Daten trainiert wird, die auf einer verzerrenden ADM-Auswahl beruhen (O'Neil 2016a: 28).

Eine solche Verzerrung wird begünstigt, wenn das Verfahren die Falsifizierung einer bestimmten Gruppe von Prognosen systematisch erschwert. Bei dem Bewährungsbeispiel ist das der Fall, wenn Richter bei hohen Risikoprognosen eher Haft als Bewährung anordnen. Fälschlicherweise als zu riskant eingeschätzte Menschen haben dann keine Möglichkeit zu beweisen, dass sie auf Bewährung nicht rückfällig geworden wären. Solche möglichen systematischen Verzerrungen müssen vor dem Einsatz eines Verfahren schon bei der Gestaltung gesucht, geprüft, diskutiert und bekämpft werden. Das Beispiel aus der Justiz zeigt, dass es dabei nicht nur um die Gestaltung eines Algorithmus oder eines Softwarepakets geht. Wäre etwa die Konsequenz eines hohen Scoringwerts für den Angeklagten nicht Haft, sondern Bewährung mit intensiver Betreuung, ließe das Verfahren möglicherweise mehr Raum für Falsifikation. Also sind auch die aus einer ADM-Prognose möglichen und die tatsächlich gezogenen Konsequenzen bei der Analyse einzubeziehen. Abhängig davon, wie ein ADM-Prozess in die Gesellschaft eingebettet ist, können die Prognosen Teilhabe einschränken. Wenn zum Beispiel qua Design Haft die einzige Konsequenz aller Risikoprognosen sein kann, betont das Risikominimierung gegenüber anderen Funktionen des Justizvollzugs, wie etwa der Resozialisierung (Christin, Rosenblat und Boyd 2015: 9).

Auch die COMPAS-Verfahren zugrunde liegende Fairness-Definition hätte vor dem Einsatz eine breite gesellschaftliche Debatte verdient. Bei der Gestaltung von ADM-Prozessen müssen Entwickler sich für eine Umsetzung

von Fairness entscheiden. Diese Operationalisierung kann in einigen Fällen zwangläufig mit einer normativen Setzung einhergehen, welche Fairnessdefinition als gerecht gewählt wird. Solchen Entscheidungen müsste eine gesellschaftliche Debatte vorausgehen, weil grundsätzliche gesellschaftliche Fragen berührt sind. Etwa bei der Prognose der Rückfallwahrscheinlichkeit von Verurteilten: Ist es fair, wenn jeder Mensch mit schwarzer Hautfarbe wahrscheinlich eine höhere Risikoprognose erhält, weil die Rückfallquote bei Menschen mit schwarzer Hautfarbe höher ist? Oder ist es fair, wenn Menschen mit weißer und schwarzer Hautfarbe, die nicht rückfällig werden, davor derselben Risikokategorie zugeordnet wurden? Beide Fairnessdefinitionen schließen einander aus. Die Debatte darüber, welche davon gerecht ist, begann in den USA erst, nachdem entsprechende ADM-Prozesse schon jahrelang bei Gericht im Einsatz waren, bis eine unabhängige Auswertung der Entscheidungen umstrittene Tendenzen offenbarte. Welchem Fairnessprinzip ein ADM-Verfahren folgen sollte, muss in solchen Fällen gesellschaftlich ausgehandelt werden. Unterbleibt eine solche Debatte wie im Fall des COMPAS-Scoring, objektiviert das Verfahren normative Festlegungen seiner wenigen Gestalter.

weitere Chancen	**weitere Risiken**
Eine algorithmenbasierte Prognose ist – anders als ein Richter – nicht tagesformabhängig (z. B. von Uhrzeit und Pausen). Eine Untersuchung von 1112 Urteilen über die Aussetzung von Strafen zu Bewährung in Israel ergab, dass die Wahrscheinlichkeit einer für den Angeklagten positiven Entscheidung am Anfang des Tages und nach Essenspausen größer ist als zu anderen Zeiten (Danziger, Levav und Avnaim-Pesso 2011: 6890).	Gerechtigkeit ist individualisiert. Urteile müssen auf dem Einzelfall und dem Handeln des Individuums basieren – nicht auf der Ähnlichkeit zu Normgruppen. Basieren Risikoprognosen nicht auf eben solchen Vergleichen? Hier kommt es darauf an, dass die Richter wirklich den Einzelfall bewerten. Es gibt Hinweise darauf, dass Menschen, wenn sie von Risikoprognosen abweichen, vor allem zulasten der Bewerteten entscheiden und trotz günstiger Prognose Haft anordnen (Steinhart 2006: 70). Hinzu kommt: Der Anwendungsbereich eingesetzter Software ist potenziell um ein Vielfaches größer als die eines Richters: Die einmal vom Entwicklerteam geformte Entscheidungslogik wird in weit mehr Fällen greifen als die Entscheidungslogik eines einzelnen Richters.
Die Inhaftierungsquote kann sinken, weil bei niedrigen Risikoprognosen Richter eher Alternativen zur Haft erwägen könnten. In Virginia steigt die Inhaftierungsrate seit Einführung von Risikoprognosen 2002 deutlich langsamer. 2014 verurteilten Richter in Virginia fast die Hälfte der Angeklagten bei Verbrechen ohne Gewaltanwendung zu Haftalternativen (wie Rehabilitationsprogrammen). Seit 2005 ist die Gefangenenzahl in Virginia um 5 Prozent gestiegen, im Jahrzehnt davor waren es 31 Prozent (Angwin et al. 2016).	Menschen empfinden softwarebasierte Prognosen als verlässlicher, objektiver und aussagekräftiger als andere Informationen zu einem Fall, einschließlich des eigenen Eindrucks (Hannah-Moffat, Maurutto und Turnbull 2009). Das kann dazu führen, dass Prognosen im Einzelfall nicht hinterfragt werden. Dabei kann die Prognose durchaus von menschlichen Fehleinschätzungen beeinflusst sein. Die Datenbasis bei Risikoprognosen kann Verzerrungen enthalten, die der Scoringwert scheinbar objektiviert: Wird zum Beispiel die Frage „Wann hatten Sie zuerst Kontakt mit der Polizei?“ einbezogen wie beim Prognoseverfahren LSI-R, verzerrt das die Risikoprognosen zulasten von Menschen aus Vierteln mit hoher Armut, Kriminalität und Polizeipräsenz (O’Neil 2016a: 27).

2.2 Sachgerechte Anwendung sichern: Individuelle Kriminalitätsprognosen

20 der 50 größten kommunalen US-Polizeibehörden nutzen Predictive Policing (Robinson und Koepke 2016: 20). Die Polizei in Chicago ordnet seit 2013 mithilfe eines ADM-Prozesses vorbestrafte Bürger aus Polizeidatenbanken einer sogenannten „Strategic Subject List“ (SSL) zu. Das Verfahren entwickelt ein Team am Illinois Institute of Technology, finanziert aus Mitteln des US-Justizministeriums.

2.2.1 Output: Software prognostiziert Opfer und Täter von Gewaltdelikten

Etwa 1400 vorbestrafte Bürger stehen auf der SSL-Liste in Chicago. Jeder erhält einen Scoringwert zwischen 1 und 500. Je höher der Wert, desto höher das Risiko, zukünftig als Täter oder Opfer in eine Schießerei oder einen Mord verwickelt zu sein (Johnson 2016: 1). So steht es in der SSL-Dienstanweisung. Doch selbst der leitende Polizeidirektor Eddie Johnson stellt öffentlich die Prognose allein als Werkzeug zur Identifizierung von Gefährdern dar, wenn er davon spricht, dass die 1400 Gelisteten für den „Großteil der Gewalt in der Stadt verantwortlich“ seien (Davey 2016). Grundsätzlich entscheiden Ermittler, wie sie die Prognosen der SSL für ihre Arbeit nutzen.

2.2.2 Datenbasis und Entscheidungslogik: Keine Transparenz

Zehn Variablen aus den Polizeidatenbanken soll der ADM-Prozess auswerten. Welche das sind und wie sie ausgewertet werden, hält die Polizei unter Verweis auf „proprietäre Technologie“ geheim. Diese Beispiele für relevante Informationen im Prozess nannte ein Polizeivertreter: Erlitt eine Person Schussverletzungen? Steigt oder fällt die Kriminalitätstrendlinie der Person? Gab es Verhaftungen wegen Waffendelikten? (a.a.O.).

2.2.3 Konsequenzen: Gefährderansprachen und Verhaftungen

Polizisten besuchten etwa 1300 Personen mit hohen Scoringwerten zu Gefährderansprachen, oft zusammen mit Sozialarbeitern, um Hilfe anzubieten (a.a.O.). Beamte können die Scoringwerte auch für Ermittlungen nutzen, jeder Polizist hat Zugang zu der Datenbank (Johnson 2016: 1). Als hinreichender Verdacht, etwa für Hausdurchsuchungen, gilt ein hoher Wert nicht. Dennoch korreliert die Wahrscheinlichkeit einer Verhaftung mit der Aufnahme in die Liste, wie die Evaluation durch die RAND Corporation ergab: „One potential reason why being placed on the list resulted in an increased chance of being arrested for a shooting is that some officers may have used the list as leads to closing shooting cases“ (Saunders, Hunt, Hollywood, Criminol und Org 2016: 1).

2.2.4 Evaluation: Kaum Prävention, mehr Verhaftungen, kein Einfluss auf Gewaltdelikte

Die RAND Corporation evaluiert das Projekt. Fazit zum ersten Einsatzjahr: Es ist keine wirksame Prävention bei den Gelisteten feststellbar, das Instrument prognostizierte zwischen März 2013 und März 2014 99 Prozent der Mordopfer nicht (Saunders 2016). Im Mai 2016 gab die Polizei an, im Jahresverlauf bis dahin seien mehr als 70 Prozent der Erschossenen und mehr als 80 Prozent der im Zusammenhang mit Schießereien Verhafteten auf der SSL gelistet gewesen (Davey 2016). Zu den Angaben für 2016 gibt es keine weiteren Details, unabhängige Untersuchungen dazu liegen nicht vor. Verhaftungen sagen nichts darüber aus, ob es sich tatsächlich um Täter handelt. Hier könnte ja auch dieser Effekt wirken: Ermittlungsdruck wird zunächst auf diejenigen ausgeübt, die der Polizei bereits bekannt sind.

> *„However, the Chicago Police failed to provide any services or programming. Instead they increased surveillance and arrests — moves that did not result in any perceptible change in gun violence during the first year of the program. (…) The names of only three of the 405 homicide victims murdered between March 2013 and March 2014 were on the Chicago police's list, while 99 percent of the homicide victims were not“ (Saunders 2016: 1).*

Auch wenn die Prognosequalität und die Präventionseffektivität rapide steigen sollten, erwartet RAND nur geringen absoluten Nutzen. Um die Mordrate in der Stadt um fünf Prozentpunkte zu senken, wären enorme Fortschritte nötig – die Prognosequalität müsste sich im Vergleich zum ersten Jahr verzehn- und die Wirksamkeit der Interventionen bei potenziellen Opfern und Tätern verfünffachen. RAND plädiert daher dafür, andere Ansätze nicht auszublenden: „And after all that improvement — here's how many lives would be saved: 21. In a city that reported 468 murders last year, that would be tremendous progress but hardly the definitive solution“ (ebd.).

2.2.5 Zur Diskussion: Es kommt auch auf die sachgerechte Einbettung an

Die Strategic Subject List wurde in Chicago nach den vorliegenden Quellen als Werkzeug für die Prävention entwickelt. Doch im tatsächlichen Einsatz wurde das Werkzeug kaum so wie geplant genutzt. Eine nennenswerte Implementierung der Prognosen für präventive Interventionen in der Polizeiarbeit konnten die evaluierenden Forscher nicht feststellen. Ihr Fazit:

> *„Overall, the observations and interview respondents indicate there was no practical direction about what to do with individuals on the SSL, little executive or administrative attention paid to the pilot, and little to no follow-up with district commanders. These findings led the research team to question whether this should be considered a prevention strategy" (Saunders et al. 2016: 10)*

Dieses Beispiel zeigt, dass die operative Anwendung und Umsetzung von Konsequenzen auch über die Wirkung eines ADM-Prozesses entscheidet. Dabei ist nicht nur die bewusste Zweckentfremdung ein Risiko (vgl. Kapitel 2.9. Zweckentfremdung verhindern), sondern auch unsachgemäße Umsetzung wie im vorliegenden Beispiel aus Chicago.

Für die eigentlich geplante Präventionsarbeit auf Basis der SSL-Prognosen fehlten in Chicago die personellen Ressourcen. Das vorhandene Personal nutzt die Prognosen offenbar stattdessen entsprechend der bestehenden Institutionslogik als Ermittlungswerkzeug. So kann die Software den Blick der Ermittler bei der Suche nach Verdächtigen auf die Personen auf der Risikoliste verengen. Solche Mechanismen bedrohen die Unschuldsvermutung und drohen, die Wirksamkeit der Polizeiarbeit zu gefährden. Die SSL ist nach öffentlicher Darstellung nicht als Werkzeug für die Fahndung nach Taten entwickelt worden. Wie gut das System für diesen Einsatz geeignet ist, müsste unabhängig evaluiert werden. Das Beispiel zeigt, dass die Güte von ADM-Prozessen auch an der operativen Einbettung in Institutionen und vor allem der tatsächlichen sachgerechten Nutzung zu messen ist.

weitere Chancen	weitere Risiken
Polizeiressourcen könnten auf Basis der Prognosen effektiver und effizienter genutzt werden.	Der Ansatz des ADM-Prozesses reduziert erfolgreiche Polizeiarbeit auf einen Aspekt: Verdächtige identifizieren. Falschprognosen des Systems wird kaum Beachtung geschenkt, was falsche Anreize bei der Anwendung setzen kann. Wie viele Menschen auf der Liste wurden zu Unrecht verdächtigt, als Gefährder angesprochen oder sogar verhaftet? Dies wird nicht evaluiert. Alternativer Indikatoren der Wirkung des Systems (z. B. Vertrauen in die Polizei in einzelnen Vierteln, exzessive Gewalt bei Einsätzen) werden nicht erfasst. Diese Faktoren beeinflussen aber die Kooperationsbereitschaft der Einwohner mit der Polizei, können so Aufklärung verbessern (The Leadership Conference on Civil and Human Rights et al. 2016: 2).
Präventionsarbeit könnte effektiver und effizienter werden und in Folge kann bestenfalls die Kriminalität zurückgehen.	Die Intransparenz der Entscheidungslogik macht eine alle Aspekte umfassende öffentliche Debatte unmöglich (a.a.O.: 1).

.

2.3 Geeignete Wirkungslogik finden: Prognose drohender Bleivergiftungen

In Chicago wurden fast 90 Prozent des Wohnungsbestandes vor 1978 erbaut – dem Jahr, in dem bleihaltige Farben in den USA verboten wurden (Potash et al. 2015: 2039). Deshalb sind Bleivergiftungen bei Kindern noch immer ein großes Problem in der Stadt: 2013 hatten zehn Prozent der unter Sechsjährigen in Chicago Bleikonzentrationen über dem vom von der US-Gesundheitsbehörde Centers for Disease Control and Prevention aufgestellten Grenzwert – das ist das Vierfache des US-Durchschnittswerts (Hawthorne 2015). Die bisherigen Maßnahmen der Stadt setzen erst an, wenn eine Bleivergiftung bei einem Kind diagnostiziert wird. Erst danach kann die Sanierung von Häusern angeordnet werden (Potash et al. 2015: 2040). Für präventive Gebäudesanierung mit öffentlichen Mitteln fehlen politische Mehrheiten (Hawthorne 2015).

Die Stadt entwickelt mit der University of Chicago eine Software, die prognostizieren soll, in welchen Gebäuden und bei welchen Kindern das Risiko einer Bleivergiftung besonders hoch ist, um frühe, zielgerichtete und daher günstige Interventionen zu ermöglichen.

2.3.1 Output: Ranking besonders gefährdeter Kinder und gefährlicher Gebäude

Die Gesundheitsbehörde Chicagos will die Software einsetzen, um anhand eines Risikorankings Gebäude und Kinder für weitere Maßnahmen zu priorisieren. Je höher das Risiko einer Bleivergiftung, desto höher sollen betroffene Gebäude und betroffene Kinder in der Rangfolge eingeordnet werden (Potash et al. 2015: 2042).

2.3.2 Datenbasis und Entscheidungslogik: Bluttests und Inspektionen

Als Datenbasis stehen den Forschern zu Verfügung: 2,5 Millionen Bluttestergebnisse auf Bleivergiftungen von etwa einer Million Kindern in Chicago zwischen 1993 und 2013 mit Datum, Identität der Testpersonen, Alter und Wohnorten. Zudem die Ergebnisse von 120.000 Hausinspektionen im selben Zeitraum mit Datum- und Ortsangaben. Die Forscher teilten die Datensätze und trainierten mehrere Klassifikationsverfahren an einem Teil der Daten, um dann Prognosen über den Zeitraum zu treffen, in dem der andere Teil der Daten liegt (a.a.O.: 2041) Die ersten in einem Peer-Review-Verfahren veröffentlichten Ergebnisse der Forscher zeigen, dass insbesondere Informationen über Alter, Inspektionsergebnisse und Zustand von Gebäuden auf Adressebene in den Trainingsdaten die Prognosequalität verbessern (a.a.O.: 2044).

2.3.3 Mögliche Konsequenzen: Untersuchungen der Hochrisikogebäude

Als mögliche Konsequenzen eines Risikorankings nennt die Behörde priorisierte Untersuchungen der Hochrisikogebäude durch Inspektoren und die folgende Renovierung bei Überschreiten von Grenzwerten. Die Forscher schlagen als denkbare Maßnahmen zudem vor: Zielgerichtete Werbung für Bluttests in Hochrisikostraßen, Veröffentlichung adressbezogener Risikoprognosen zur Orientierung für Mieter, zielgerichtete Ansprache von Vermietern auf Basis der Prognosen (a.a.O.: 2046).

2.3.4 Evaluation: Läuft

Die Gesundheitsbehörde Chicagos validiert das Modell derzeit (Chicago Department of Public Health 2016: 3), die Anwendung scheint sich auf adressbezogene Risikoprognosen zu fokussieren.

2.3.5 Zur Diskussion: Effizienzgewinne sind nicht immer eine hinreichende Wirkungslogik

Sollten die Prognosen sich als treffsicher erweisen, könnte die Stadt Chicago zielgerichtet Eltern von Kindern mit dem höchsten Risiko für Bleivergiftungen ansprechen und Hochrisikogebäude als erste renovieren. Das wäre eine Verbesserung gegenüber dem Status quo, bei dem zu geringe Mittel zu breit gestreut werden. Möglich macht das ein grundsätzlicher Vorteil maschineller Entscheidungen: Algorithmische Verfahren können weit mehr Faktoren und Daten auswerten als Menschen.

Dieser Vorteil garantiert aber nicht allein, dass mehr Teilhabechancen für alle entstehen. Ein grundsätzliches Risiko, das selbst bei treffsicheren Prognosen weiter bestehen kann: Der durch Algorithmen ermöglichte zielgerichtete Ressourceneinsatz kann die Frage überlagern, ob der Art und dem Umfang der Maßnahmen eine

geeignete Wirkungslogik zugrunde liegt. Um diese zu entwickeln, müssen die übergeordneten Ziele und die Handlungsoptionen zum Erreichen transparent gemacht werden.

Bei dem Fallbeispiel fällt auf, dass der Einsatz des ADM-Prozess losgelöst davon diskutiert wird, welche Mittel für die Untersuchung und Renovierung von Gebäuden verfügbar sind. Was geschieht nach der Prognose eines erhöhten Risikos auf Bleivergiftung? Die Stadt Chicago beschäftigte im Jahr 2015 elf Inspektoren für die Untersuchungen von Häusern auf Bleibelastung und drei Krankenschwestern. Das ist nur noch knapp ein Viertel des Personals, das für diese Belange 2010 zu Verfügung stand (Hawthorne 2015). Womöglich kompensieren etwaige Effizienzgewinne durch den Einsatz von ADM-Prozessen also lediglich einen durch Einsparungen geschaffenen Mangel – wenn überhaupt. Vielleicht sollten für den Schutz von Kindern vor Bleivergiftungen insgesamt mehr Mittel aufgewendet werden. Vielleicht reicht auch der bisherige Wirkmechanismus nicht aus oder ist dem angestrebten gesellschaftlichen Ziel nicht angemessen: Vielleicht sollte der Fokus darauf liegen, Bleivergiftungen zu vermeiden, statt „nur“ Eltern zu Bluttests animieren, um Bleivergiftungen bei Kindern diagnostizieren zu können. Solche Fragen kann die Gestaltung eines ADM-Prozesses nicht lösen. Aber diese Fragen bestimmten den Rahmen und die Ziele der Gestaltung eines ADM-Prozesses.

2.4 Konzepte korrekt messbar machen: Armutsverteilung vorhersagen

Um Entwicklungshilfe zielgerichtet einzusetzen und die Wirkung von Maßnahmen zu bewerten, benötigt man aktuelle Informationen über die lokale Armutsverteilung. Um eine neue Datenbasis mit größerer Varianz zu erschließen, haben Forscher ein künstliches neuronales Netz darauf trainiert, in Satellitenfotos bei Tageslicht Landschaftsmerkmale zu erkennen, die mit extremer Armut zusammenhängen. Die Ergebnisse erschienen im August 2016 in Science (Jean et al. 2016), im praktischen Einsatz ist das Verfahren noch nicht.

2.4.1 Output: Ausgaben und Vermögen auf Dorfebene

Die Software prognostiziert für geographische Cluster auf Dorfebene in Nigeria, Tansania, Uganda, Ruanda und Malawi die täglichen Pro-Kopf-Ausgaben nach Weltbank-Definition und das Haushaltsvermögen nach der in der US-Entwicklungshilfe genutzten Definition des Demographic and Health Survey Program (2014).

2.4.2 Datenbasis und Entscheidungslogik: Flächendeckende Satellitenfotos und Umfragen

Daten zur Armutsverteilung könnten Umfragen zu Kaufkraft und Vermögen liefern. Solche Umfragen in ländlichen Regionen Afrikas sind aufwendig, teuer und daher selten: Zwischen 2000 und 2010 haben 39 von 59 Staaten in Afrika weniger als zwei solcher Umfragen durchgeführt (Patel 2016). Daher suchen Forscher andere Datenquellen, die Aussagen über die Armutsverteilung auf Dorfebene ermöglichen. Daten zur Mobilfunknutzung haben eine gewisse Aussagekraft, sind aber nicht öffentlich verfügbar. Satellitenaufnahmen bei Nacht sind öffentlich verfügbar, allerdings ist die Aussagekraft in Regionen geringer, in denen viele Menschen in extremer Armut (nach Definition der Weltbank 2015) leben: Wo extreme Armut herrscht, ist es nachts fast durchweg dunkel, die Abstufungen sind sehr gering (Jean et al. 2016: 790).

Deshalb nutzt das Forscherteam des Sustainability and Artificial Intelligence Lab der Stanford University Satellitenfotos bei Tag und Nacht sowie vorliegende Umfrageergebnisse zu Pro-Kopf-Ausgaben und Haushaltsvermögen. Anhand dieser Daten trainiert es in mehreren Schritten künstliche neuronale Netzwerke. Im ersten Schritt arbeitete ein neuronales Netzwerk Eigenschaften von Tagesaufnahmen heraus, die mit den Lichtunterschieden in den Nachtaufnahmen zusammenhängen. Einige dieser Merkmale sind auch für Menschen erkennbar, zum Beispiel Straßen, städtische Siedlungsgebiete, Ackerflächen (Horton 2016). Im zweiten Schritt trainierten die Forscher ein künstliches neuronales Netzwerk darauf, zu erkennen, welche der herausgearbeiteten Eigenschaften in Tageslichtaufnahmen mit der für diese Region in Umfrageergebnissen festgestellten Armutsverteilung zusammenhängen. Die Software hat zum Beispiel herausgearbeitet, dass die Materialbeschaffenheit (Metall, Stroh, Erde, Gras) von Dächern mit den Pro-Kopf-Ausgaben zusammenhängt (Jean et al. 2016: 791).

2.4.3 Evaluation: Bessere Ergebnisse als mit Mobilfunkdaten

Das Verfahren liefert bessere Ergebnisse als auf Mobilfunknutzung beruhende Methoden. Der Vergleich zeigt, dass die Prognosequalität der Haushaltsvermögen auf Dorfebene in Ruanda besser ist (Korrelationskoeffizient 0,62 bei der Mobilfunkmethode vs. 0,75 bei der Mustererkennung in Tageslichtsatellitenfotos) (a.a.O.: 792). Die Studie zeigt zudem, dass die an den Umfragedaten eines Staates trainierten Modelle sich in anderen Staaten anwenden lassen:

> *„Pooled models trained on all four consumption surveys or all five asset surveys very nearly approach the predictive power of in-country models in almost all countries for both outcomes. These results indicate that, at least for our sample of countries, common determinants of livelihoods are revealed in imagery, and these commonalities can be leveraged to estimate consumption and asset outcomes with reasonable accuracy in countries where survey outcomes are unobserved“ (a.a.O.: 794).*

2.4.4 Mögliche Konsequenzen: Zielgerichteter Einsatz von Hilfsmaßnahmen

Derzeit ist unbekannt, welche Konsequenzen Hilfsorganisationen auf Basis der Prognosen umsetzen würden. Denkbar sind positive Folgen für die von den Prognosen betroffenen Menschen (Ausbau der Hilfsmaßnahmen).

2.4.5 Zur Diskussion: Es braucht öffentlichen Diskurs über Operationalisierungen

Wenn algorithmische Verfahren auf Basis von vorhandener Satellitenfotografie die Armutsverteilung zuverlässig prognostizieren, könnten Maßnahmen der Entwicklungshilfe erheblich günstiger, aktueller und vor allem bedürfnisgerechter werden. Grund dafür ist eine gemeinsame Eigenschaft aller ADM-Verfahren: Eine einmal entwickelte Entscheidungslogik lässt sich zu vergleichsweise geringen Kosten auf beliebig viele Fälle anwenden. Sie lässt sich zum Teil sogar unter anderen Rahmenbedingungen anwenden, wie etwa im vorliegenden Fallbeispiel auf andere Staaten.

Möglich sind diese Fortschritte, weil der ADM-Prozess auf die Prognose eindeutiger Zielgrößen, wie hier die Pro-Kopf-Ausgaben, optimiert werden kann. Diese Zielgröße ist in der Entwicklungshilfe schon lange im Einsatz, für ihre Aussagekraft stehen Institutionen wie die Weltbank. Neu ist der Prognoseansatz, nicht der Messwert. Es ist von Vorteil für den ADM-Einsatz, wenn dessen Zweck und Operationalisierung zuvor Gegenstand eines (fach-)öffentlichen Diskurses war.

Da die Entwicklung der Instrumente erst am Anfang steht, kann über die Folgen im Einsatz nur spekuliert werden. Die Qualität der Prognosen müsste durch Abgleich mit Umfrageergebnissen in Stichproben geprüft werden. Das muss im Praxiseinsatz geschehen. Sollte der ADM-Prozess in der Praxis für die örtliche Verteilung von Hilfsmaßnahmen genutzt werden, sind auch negative Folgen denkbar. So könnten Hilfsmaßnahmen an den im Landesvergleich bessergestellten Orten verringert werden, auch wenn dort nach internationalen Standards die Armut problematisch hoch ist.

2.5 Umfassende Evaluation gewährleisten: Automatische Gesichtserkennung

Die US-Bundespolizei FBI betreibt seit 2008 ein System, das per Gesichtserkennung Fotos unbekannter Personen analysiert und Übereinstimmungen in verschiedenen Datenbanken mit gut 400 Millionen Porträts von US-Bürgern und Ausländern (z. B. aus Visaanträgen) sucht. Der US-Rechnungshof kritisiert, dass Zuverlässigkeit und Fehlerquote des gesamten Systems nie getestet worden sind (United States Government Accountability Office 2016).

2.5.1 Output: Bis zu 50 Vorschläge für die gesuchte Person

Ermittlungsbehörden, denen zu einer konkreten Straftat Bilder von Verdächtigen vorliegen, können das FBI bitten, diese Bilder mit seiner bestehenden Datenbank abzugleichen. Ziel einer solchen Abfrage ist entweder die Bestätigung eines bereits bestehenden Verdachts oder die Bitte um eine Liste möglicher Verdächtiger, deren hinterlegte biometrische Kriterien zu denen des gesuchten Verdächtigen passen. Abfragen können grundsätzlich nur über das FBI gestellt werden. Seit 2011 können im Rahmen eines Pilotprojektes jedoch auch Ermittlungsbehörden aus sieben einzelnen Bundesstaaten direkt auf die Datenbank zugreifen – zwischen 2011 und 2015 führten sie mehr als 20.000 Suchen durch (interessanterweise rangiert die Anzahl der Anfragen pro Bundesstaat von 20 bis 14.000). Wird eine Suchabfrage gestellt, vergleicht der Algorithmus das Foto der gesuchten Person mit den abgelegten biometrischen Informationen und generiert eine Liste von zwei bis 50 Vorschlägen, wer die abgebildete Person sein könnte. Diese Informationen werden im Falle eines indirekten Zugriffs von sogenannten „biometrischen Analysten“ manuell überprüft und auf einen bis zwei Kandidaten reduziert, die dann vom FBI an die anfragende Organisation (z. B. lokale Polizei) weitergegeben werden können. Beim FBI arbeiteten 2015 29 solcher Analysten. Im Falle eines direkten Zugriffs durch eine bundesstaatliche Ermittlungsbehörde ist dieser Zwischenschritt nicht gewährleistet. Das Verfahren wird nicht nur für die Ermittlungen bei Gewaltverbrechen verwendet, sondern auch bei Diebstahl oder Versicherungsbetrug (a.a.O.).Ermittler entscheiden auf Basis der vorgelegten Verdachtsliste, ob und gegen welche der angegebenen Personen sie weitere Schritte einleiten.

2.5.2 Datenbasis und Entscheidungslogik: „Criminal identities“ vs. „civil identities“

Die Datenbank wird durch freiwillige Einsendungen der verschiedenen US-Behörden gespeist und umfasst derzeit mehr als 30 Millionen Bilder, die nach Vorbestrafungen in „criminal identities“ und „civil identities“ unterschieden werden: Erstere sind Bilder, die im Rahmen von Verhaftungen, Verurteilungen oder Gefängnisstrafen erstellt wurden. Mehr als 80 Prozent der Aufnahmen gehören zu dieser Kategorie. „Civil identities“ stammen aus Personalakten, Militärdienst, Freiwilligendienst oder Einwanderungspapieren. Jedes Bild wird mit einem vollständigen Set an Fingerabdrücken verbunden, sodass Duplikate automatisch verknüpft werden (auch zwischen den Identitäten). Einmal eingestellte Bilder können aus der Datenbank nur durch die einsendende Organisation oder per Gerichtsbeschluss entfernt werden. Über eine spezielle Abteilung (FACE) kann das FBI zudem auf andere staatliche Datenbanken zugreifen (bspw. Bilder aus Führerscheinen und Visa-Anträgen). Inklusive dieser externen Datenbanken beläuft sich die Zahl der verfügbaren Bilder auf mehr als 411 Millionen – und betrifft rund 64 Millionen Amerikaner (Garvie, Bedoya und Frankle 2016).

Die Entscheidungslogik besteht aus zwei Stufen: Eingesendete Bilder werden auf biometrische Kriterien hin analysiert und in der Datenbank abgelegt. Zivile Datensätze kann nur vom FBI durchsucht werden, während Anfragen zu den Datensätzen vorbestrafter Personen allen Strafverfolgungsbehörden offenstehen, die auch Bilder einsenden. Sofern allerdings zivile mit kriminellen Datensätzen verbunden sind, werden beide angezeigt. Die Liste mit den vielversprechendsten Matches wird dann im Falle einer indirekten Anfrage zur Human Analysis weitergeleitet (s. „Output“), bei direkten Anfragen an die das System nutzende Ermittlungsbehörde (United States Government Accountability Office 2016). Über die verwendeten Kriterien des Algorithmus ist nichts bekannt.

2.5.3 Konsequenzen: 64 Millionen Menschen in einer fortwährenden Gegenüberstellung

Im Falle einer fälschlichen Zuordnung einer Person aus der Datenbank zu einem Fahndungsfoto kann die vorgeschlagene Person zu Unrecht verdächtigt werden. Je nach Bundesstaat gilt eine positive Übereinstimmung sogar

als Beweismittel vor Gericht. Somit spielt der Algorithmus eine Rolle bei der Entscheidung über Freiheit oder Gefängnisstrafe von Bürgern. Aber schon „nur“ auf der Ergebnisliste zu stehen, hat diverse Konsequenzen für die Betroffenen: Die so erhaltenen Hinweise können als Grundlage für Hausdurchsuchungen, Datenabfragen bei Internetprovidern und Banken sowie Verhaftungen genutzt werden.

Hinzu kommen ethnische Diskriminierungen: So verfügt die Datenbank der Vorbestraften über mehr Bilder von Menschen mit weißer als mit schwarzer Hautfarbe. Das führt dazu, dass der Algorithmus in ersterer Population mit höherer Wahrscheinlichkeit Übereinstimmungen findet als in letzterer.

Nicht zuletzt bedeutet die Verwendung von Gesichtserkennungsalgorithmen juristisch eine Einschränkung der Unschuldsvermutung auf zwei Ebenen: Zum einen reicht der Vorschlag des Algorithmus als Anfangsverdacht aus, um Ermittlungen gegen die betroffene Person einzuleiten. Zum anderen kann eine positive Übereinstimmung in manchen Bundesstaaten als Schuldbeweis gewertet werden. Die Falsifikationsrate des verwendeten Algorithmus beschreibt die Wahrscheinlichkeit, unschuldige Bürger auf Basis eines womöglich falsch zugeordneten Datenbankeintrages zu verdächtigen oder sogar zu verhaften. Daher sollte bereits vor Implementierung eines solchen Algorithmus eine gesellschaftliche Debatte über den akzeptablen Grenzwert dieser Rate geführt werden. Es ist eine grundsätzliche Frage der Fairness, die aufgrund der Eigenarten von ADM-Prozessen früh zu operationalisieren ist: Will eine Gesellschaft zu Unrecht beschuldigte Bürger in Kauf nehmen? Wie viele? Und wie minimiert man das so verursachte Leid?

2.5.4 Evaluation: Technisch unzureichende Evaluierung, fehlender juristischer Rahmen

Verschiedene Kontrollsysteme wurden bereits vor dem Einsatz des Next Generation Identification-Interstate Photo System (NGI-IPS) aufgestellt, um die ethische Unbedenklichkeit einer verwendeten Technologie sicherzustellen: So fordert das Department of Justice (DOJ) für jedwede Technologie, die Bürgerdaten sammelt, ein „Privacy Impact Assessment“ (PIA) ein. Für das verwendete NGI-IPS wurde diese jedoch nur für das ursprünglich eingeführte, funktional sehr viel weniger umfangreiche System 2011 erstellt. Die seitdem erfolgten Updates und Erweiterungen wurden erst im September 2015 überprüft.

Zudem evaluierte das FBI intern den Erfolg des neuen Algorithmus: Dazu sollten sowohl die Erkennungsrate (Wahrscheinlichkeit, dass die gesuchte Person unter den 50 vorgeschlagenen Übereinstimmungen ist) als auch die Fehlidentifikationsrate (Wahrscheinlichkeit, dass jemand zu Unrecht als Übereinstimmung vorgeschlagen wird) überprüft werden. Für ersteren Test wurde eine Trefferquote von 85 Prozent als akzeptabel festgelegt (dies entspricht allein bei den extern veranlassten Suchen immer noch 3000 ergebnislosen Listen). Sofern die gesuchte Person also in der Datenbank vorkommt, sollte sie in 85 Prozent der Fälle auf einer Liste mit 50 Vorschlägen vertreten sein. Dieser Zielwert wurde im Rahmen der Evaluation mit einem tatsächlichen Ergebnis von 86 Prozent der Fälle erreicht. Getestet wurden allerdings nur Listen mit 50 Vorschlägen – möglich sind aber Listen von zwei bis 50 Vorschlägen, mit einer Standardeinstellung von 20 potenziellen Übereinstimmungen. Für diese kleineren Listen liegt keine Evaluation vor.

Die Fehlidentifikationsrate wurde gar nicht getestet. Das FBI argumentierte hier, dass die Listen nur Vorschläge enthielten und somit keine „positives“ seien. Sowohl der Bericht des United States Government Accountability Office (2016) als auch Garvie, Bedoya und Frankle (2016) weisen allerdings darauf hin, dass allein das Verdachtsmoment eine Abkehr vom Unschuldsprinzip darstellt – Garvie Bedoya und Frankle bezeichnen den Vorgang als „perpetual line-up“.

Unabhängig von diesem konkreten Beispiel ist sich die Wissenschaft auch über die Effektivität algorithmischer Gesichtserkennung generell uneins (Revell 2016) – zu viele Faktoren bestimmen die tatsächliche Erfolgsrate. Als ein Beispiel seien die Lichtverhältnisse genannt: Im Rahmen eines Experiments in Mainz variierte die Trefferquote des zur Gesichtserkennung in Echtzeit in einer U-Bahn-Station verwendeten Algorithmus zwischen 60 Prozent am Tag und zehn bis 20 Prozent bei Nacht (Garvie, Bedoya und Frankle 2016). Weitere erschwerende Faktoren reichen vom Winkel, in dem die Kamera auf das Gesicht sieht, der Auflösung der Kamera selbst oder der Qualität des Vergleichsfotos bis hin zu plastischer Chirurgie, Make-up und Alterungsprozessen.

Zudem stellt sich die Frage nach der Sicherung eines solchen Systems gegen unzulässigen oder unautorisierten Zugriff. Hier reichen die Szenarien von Hackerangriffen bis hin zu autorisierten Nutzern, die jedoch unzulässige Suchen durchführen (bspw. nach Verwandten) (Garvie, Bedoya und Frankle 2016). Auch hier bemängeln die Evaluatoren die nachlässigen Strukturen, mit denen das System derzeit vor solchem Missbrauch geschützt werden: So regulieren beispielsweise nur fünf Bundesstaaten die polizeiliche Nutzung von Gesichtserkennungsalgorithmen überhaupt. Bezüglich der Erkennungsrate, ab der eine Verwendung von Systemen zulässig ist, gibt es in keinem Bundesstaat Vorgaben. Und auch zur Schwere der Straftat, ab der eine Verwendung des Systems zulässig ist, variieren die Vorschriften von Bundesstaat zu Bundesstaat (a.a.O.).

Auch die Entwicklung des Algorithmus an den eigentlich vorgesehenen Kontrollpunkten wirft grundlegenden Fragen auf: Wie müsste ein Kontrollsystem aussehen, das mit dem Tempo eines sich ständig wandelnden Algorithmus mithalten und die technische und juristische Komplexität seiner Anwendung sowie die entsprechende gesellschaftliche Debatte nicht nur im Nachhinein abbilden, sondern auch präventiv steuern kann?

2.5.5 Diskussion: Umfassende Evaluation dauert an und analysiert mittelbare Folgen

Das Fallbeispiel zeigt: Die Skalierbarkeit maschineller Entscheidungen kann schnell zu Einsatzszenarien führen, deren gesellschaftliche Angemessenheit und Folgen nicht debattiert worden sind. ADM-Prozesse ermöglichen Abfragen in einer Breite und Häufigkeit, die mit analogen Mitteln nicht möglich war. Viele Polizeistellen können auf die FBI-Datenbank zugreifen. Die Datenbank verknüpft wiederum eine Vielzahl von Quellen. Diese Vernetzung und der geringe Aufwand algorithmischer Gesichtserkennung könnten dazu führen, dass

- sie für Bagatelldelikte eingesetzt wird.
- es wegen der steigenden Menge an Anfragen zu absolut mehr Fehlern kommt.
- für einige Menschen das Risiko von Fehlidentifikationen steigt, weil ihre Porträts wegen systematischer Verzerrungen in der Datenbasis enthalten sind. Zum Beispiel, weil in Vierteln mit hoher Armut die Wahrscheinlichkeit von Polizeikontrollen, Zufallsfunden und Polizeifotos in Folge höher ist.

Die Folgen dieser neuen Qualität der Gesichtserkennung sind bei dem Fallbeispiel nicht ausreichend evaluiert worden. Weder vor Einführung noch im Nachhinein wurde die Angemessenheit der fortwährenden automatischen Gesichtserkennung untersucht und diskutiert. Zu einer umfassenden Evaluation gehören auch die Abschätzung mittelbarer Folgen und die fortwährende Analyse der tatsächlichen Anwendung im Einsatz.

weitere Chancen	weitere Risiken
Erhöhung der Wahrscheinlichkeit, Straftäter zu fassen: Es gibt Beispiele für erfolgreich identifizierte Kriminelle, die sich vor Einführung der standardisierten Erkennung jahrzehntelang erfolgreich einem polizeilichen Zugriff entziehen konnten.	Der Gesichtserkennung durch ADM-Prozesse können sich US-Amerikaner nicht entziehen. Garvie, Bedoya und Frankle zufolge wurde statistisch gesehen jeder zweite Bürger in den USA bereits ohne sein Wissen einer algorithmischen Gesichtserkennung unterzogen.
Erhöhung der Ressourceneffizienz bei der Ermittlung im Rahmen von Straftaten: Das Durchsuchen einer zentralen Datenbank mit mehr als 400 Millionen Bildern nach Übereinstimmungen wäre ohne Einsatz eines Algorithmus nur in ausgewählten Einzelfällen zu rechtfertigen. Der Einsatz eines Algorithmus erlaubt den Zugriff auch bei geringfügigeren Verdachtsmomenten.	Eine Fehlidentifikation kann zur Falschverdächtigung Unschuldiger führen. Folgen dieser Falschverdächtigung können neben der Durchsuchung und Überwachung privater Datenströme auch eine Untersuchungshaft oder sogar eine Verurteilung sein.

2.5.6 Situation und Relevanz in Deutschland

Automatisierte Gesichtserkennung ist in Deutschland beispielsweise an sieben Flughäfen im Einsatz als Teil des automatisierten Grenzkontrollsystem EasyPASS (Bundespolizei 2015). Gemeinsam mit der Deutschen Bahn entwickelte das Bundesinnenministerium 2016 ein Konzept zur Überwachung von Bahnhöfen mittels

Gesichtserkennung, das bereits in 20 Bahnhöfen pilotiert wurde (Plass-Fleßenkämper 2016). Innenminister Thomas de Maizière forderte 2016 die Einführung von Systemen zur Gesichtserkennung an allen deutschen Bahnhöfen und Flughäfen („Terrorbekämpfung“ 2016). Im Februar 2017 kündigte die Deutsche Bahn an, künftig am Berliner Südkreuz intelligente Videoüberwachung mit Gesichtserkennungsfunktion zu testen: „Diese Kamera ist ein kleines Wunderding: Sie soll durch eine Gesichtserkennung Menschen herausfiltern, die auf einer Liste von Verdächtigen gespeichert sind. Zudem soll sie abgestellte Gegenstände, etwa Koffer oder Pakete, die längere Zeit nicht bewegt wurden, registrieren. Und auch das typische Verhalten von Taschendieben soll sie erkennen“ (Kurpjuweit 2017).

2.6 Vielfalt von ADM-Prozessen sichern: Bewerbervorauswahl per Online-Persönlichkeitstest

In Großbritannien und den Vereinigten Staaten werden 60 bis 70 Prozent der Bewerber automatisierten Auswahlverfahren und Tests unterzogen. Eine wichtige Rolle spielen dabei Persönlichkeitstests. Online ist der Einsatz erheblich günstiger als in persona. Deshalb nutzen Agenturen und Arbeitgeber automatisiert ausgewertete Online-Persönlichkeitstest immer häufiger und frühzeitiger im Auswahlverfahren – zunehmend auch für durchschnittlich und unterdurchschnittlich bezahlte Stellen im Dienstleistungssektor (Weber und Dwoskin 2014).

2.6.1 Output: Viele Bewerbungen sieht kein Mensch

Die automatisierten Verfahren dienen der Vorauswahl von Bewerbungen. Ein Teil wird sofort auf Basis der Onlinetests abgelehnt, noch bevor ein Mensch die Bewerbungen sieht. Wie hoch der Anteil dieser Ablehnungen ist, kann der Arbeitgeber bestimmen. Ein Testanbieter spricht von 30 Prozent automatisierter Ablehnungen (a.a.O.). Die Informatikerin Cathy O'Neil (2016a: 105) gibt an, dass 72 Prozent der Bewerbungen in den USA nur maschinell ausgewertet werden.

Nach den wenigen vorliegenden Informationen scheint die Vorauswahl ohne menschlichen Entscheider abzulaufen.

2.6.2 Datenbasis und Entscheidungslogik: Onlinefragebögen zur Persönlichkeit

Über die Entscheidungslogik der Verfahren ist nur wenig öffentlich bekannt. In den USA wehren sich Anbieter wie Kronos juristisch gegen Auskunftsersuche von US-Bundesbehörden.

Die Onlinetests enthalten Skalenfragen wie „Im Verlauf eines Tages kann ich viele Stimmungswechsel erfahren", „Wenn etwas sehr Schlimmes passiert, brauche ich einige Zeit, damit ich mich wieder glücklich fühle" (Weber und Dwoskin 2014). Diese Fragen zielen offenkundig darauf ab, Bewerber nach dem Fünf-Faktoren-Persönlichkeitsmodell zu bewerten. „People the system classified as ‚creative types' tended to stay longer at the job, while those who scored high on ‚inquisitiveness' were more likely to set their questioning minds toward other opportunities" (O'Neil 2016a: 109).

Einige Tests fragen ab, wie lange der Kandidat die Anfahrt zum neuen Arbeitsort einschätzt. Diese Informationen nutzte ein Dienstleister von Xerox Services (US-Callcenter-Betreiber, der 30.000 Bewerber jährlich einstellt) zur automatisierten Aussonderung von Kandidaten: Wer zu lange Anfahrtswege hatte, wurde abgelehnt, weil Mitarbeiter mit langen Wegen statistisch eher kündigen als andere. Xerox Services strich dieses Kriterium, weil es systematisch Menschen aus ärmeren Vierteln mit vorrangig schwarzer Bevölkerung diskriminieren könnte, die sich Wohnungen in der Nähe des Unternehmens nicht leisten können. Es ist möglich, dass Gerichte diese Praxis als verbotene Diskriminierung nach Hautfarbe werten würden, sollte jemand klagen (Weber und Dwoskin 2014).

2.6.3 Konsequenzen: Es geht nicht nur um *einen* Job, sondern um Zugang zum Arbeitsmarkt

Zur automatisierten Bewerberauswahl greifen viele Arbeitgeber in den USA auf die Software weniger Dienstleister zu. Der breite Einsatz führt dazu, dass bei unterdurchschnittlich bezahlten Tätigkeiten im Dienstleistungsbereich die Software zum Torwächter nicht für einen, sondern für die Mehrheit potenzieller Stellen wird. Welche Konsequenzen das haben kann, zeigt der Fall von Kyle Behm. Der Ingenieurstudent war nach erfolgreicher psychiatrischer Behandlung wegen einer bipolaren Störung wieder an der Universität und suchte einen Nebenjob. Er hatte zuvor in Supermärkten gearbeitet. Nun wurde er bei sieben potenziellen Arbeitgebern nach ähnlichen Onlinetests für Mindestlohn-Stellen in Fastfoodläden, Bau- und Supermärkten abgelehnt (O'Neil 2016c: 1). Behms Vater kontaktierte die Unternehmen, die Mehrheit bot nach Prüfung einen passenden Job an, sollte Behm im Gegenzug auf eine Klage verzichten. Behm reichte eine Beschwerde bei der US-Bundesbehörde Equal Employment Opportunity Commission (EEOC) ein. Die Untersuchung zum Einsatz von Persönlichkeitstests läuft.

Für bestimmte Gruppen können die automatisierten Verfahren in dieser Form den Zugang zum Arbeitsmarkt insgesamt erschweren. Eine frühere Form von Diskriminierung (z. B. nach ausländisch/dunkelhäutig klingenden Namen) kann durch eine andere ersetzt werden. Da die Technik vor allem im Niedriglohnsektor zum Einsatz kommt, dürften Unternehmen wenig geneigt sein, ausgehend von Einzelfällen in die Verbesserung und Prüfung der Systeme zu investieren. Die Systeme müssen nicht alle Besten finden, sondern nur effizienter sein als das vorherige Auswahlsystem. Investitionen in die Kalibrierung der Systeme, die fortwährende Prüfung und Aktualisierung der Entscheidungslogiken und des Datenbestandes werden auf diesem Einsatzgebiet nie so hoch sein wie bei Stellen mit geringem Angebot an Arbeitskräften, hohem Bedarf und entsprechend hohen Gehältern. Zudem gibt es kaum Möglichkeiten, die Prognose dieses ADM-Prozesses so zu falsifizieren, dass das System aus dem Fehler lernt: Auch wenn Kyle Behm nach sieben Ablehnungen andernorts einen Job bei McDonald's erhält und sich binnen zwei Jahren zum Standortmanager hocharbeitet, wird keines der sieben anderen Unternehmen den genutzten Persönlichkeitstest prüfen, um herauszufinden, warum die Prognose falsch lag (O'Neil 2016c: 4).

2.6.4 Evaluation: Es existieren keine unabhängigen Analysen

Die Wirksamkeit der automatisierten Auswahlverfahren wurde bislang nicht unabhängig getestet. Einige Unternehmen wie Xerox Services berichten von Erfolgen: Der Arbeitskräfteabgang sei an einigen Standorten um ein Fünftel gesunken, es seien auch Menschen eingestellt worden, die allein auf Basis ihres Lebenslaufs keine Chance gehabt hätten, sagte die Personalchefin von Xerox Services 2014. Eine systematische Evaluation scheint aber nicht Grundlage des Urteils zu sein: „‚I don't know why this works', admits Ms Morse, ‚I just know it works'" (Smedley 2014: 1). Diese Aussagen sind aus zwei Gründen vorsichtig zu bewerten: Es ist unwahrscheinlich, dass Unternehmen Misserfolge mit automatisierten Verfahren veröffentlichen. Und die Kennzahlen des Unternehmens erfassen nicht die Wirkung des Verfahrens auf abgelehnte Bewerber, etwa die Frage, ob es zu systematischen Verzerrungen kommt.

Der US-Fachverband Society for Human Resource Management gab bei einer Anhörung der Equal Employment Opportunity Commission an, es gäbe keine empirischen Belege für Validität oder nachteilige Effekte der Testverfahren (Dunleavy 2016). Eine Metastudie von 7000 Veröffentlichungen ergab, dass Persönlichkeitstest nur sehr geringe Aussagekraft über zukünftige Leistungen am Arbeitsplatz haben (Morgeson et al. 2007).

2.6.5 Zur Diskussion: Je breiter der ADM-Einsatz, desto wirkmächtiger sind auch seltene Fehler

Der Fall von Kyle Behm zeigt ein mögliches strukturelles Problem von ADM-Prozessen: Menschen wenden Leitlinien etwa zur Bewerberauswahl dezentral an – oft vielfältig und unterschiedlich. Software hingegen wendet den festgelegten Prozess in jedem Einzelfall gleich an. Bei einem dezentralen Prozess hätte einer der Entscheider Kyle Behm vielleicht eine Chance geben. Diese Zentralisierung einer Entscheidungslogik wiegt umso schwerer, je mehr Institutionen dieselben ADM-Prozesse einsetzen. Durch die in den USA breite Anwendung von bestimmten ADM-Prozessen weniger Dienstleister dominieren nach den vorliegenden Quellen in bestimmten Sektoren wenige Verfahren. Zum Beispiel bei der Vorauswahl für gering bezahlte Dienstleistungsjobs. Das kann dazu führen, dass ohnehin marginalisierte Gruppen sich den automatisierten Verfahren kaum entziehen können.

Rechtsmittel scheinen bei diesen Fällen derzeit ein schwaches Korrektiv zu ein. Denn die Intransparenz der Verfahren und Entscheidungslogiken erschwert es Bewerteten, einen Ansatzpunkt für eine Beschwerde oder gar Klage zu entdecken. Zudem sind mögliche Kläger im Niedriglohnsektor finanzschwach. Jeder abgewiesene Bewerber wird abwägen, ob es die Mühe lohnt, juristische Schritte einzuleiten statt alle Ressourcen in die weitere Jobsuche zu stecken.

Hinzu kommt der auch in Kapitel 2.1 wirkende Effekt strukturell verzerrter Falsifikation: Sollte einer der zur Bewerberauswahl genutzten ADM-Prozesse systematisch geeignete Kandidaten aussortieren, gibt es beim vorliegenden Aufbau keine Möglichkeit, dass das System diesen Fehler erkennt und daraus lernt. Die Kosten einer ungerechtfertigten Ablehnung sind bei Jobs im Niedriglohnsektor für die ablehnenden Unternehmen gering und der Anreiz für die dahingehende Optimierung des Verfahrens entsprechend niedrig. Wenn es keine Möglichkeit der Falsifikation der Prognosen gibt, fällt wie beim breiten Einsatz einzelner Verfahren im Niedriglohnsektor ein wichtiges

Qualitätselement für ADM-Prozesse weg (O'Neil 2016c: 4). Abhilfe kann die Vielfalt unterschiedlicher Verfahren schaffen.

weitere Chancen	weitere Risiken
Formale Qualifikation kann bei automatisierten Testverfahren weniger zählen als bei herkömmlichen. Das eröffnet Chancen für bisher benachteiligte Gruppen wie Langzeitarbeitslose oder Niedrigqualifizierte. Kompetenzen zählen mehr als Kreditpunkte, Arbeitsmarktbedarfe mehr als Abschlüsse.	Es gibt Hinweise, dass einige der Verfahren bestimmte Gruppen benachteiligen. Zum Beispiel Menschen aus Wohngegenden mit niedrigen Einkommen oder Menschen mit psychischen Erkrankungen, die sie zwar beim Persönlichkeitstest beeinträchtigen, aber nicht unbedingt bei der Tätigkeit, für die rekrutiert wird.
Diskriminierung auf Basis von Faktoren wie Geschlecht, fremd klingenden Namen, Bewerberfotos oder offen kommunizierten Behinderungen kann sinken. Zur Veranschaulichung des Status quo: „Um eine Einladung zum Vorstellungsgespräch zu erhalten, muss ein Kandidat mit einem deutschen Namen durchschnittlich fünf Bewerbungen schreiben, ein Mitbewerber mit einem türkischen Namen hingegen sieben" (Schneider, Yemane und Weinmann 2014: 4).	Wenn die ADM-Prozesse an Daten trainiert werden, die aus dem bisherigen Auswahlprozess resultieren, kann das bisherige Diskriminierung fortschreiben (Trindel 2016).

2.6.6 Situation und Relevanz in Deutschland

In Deutschland setzten einer Umfrage aus dem Jahr 2016 zufolge „6,0 Prozent der 1000 größten deutschen Unternehmen aktuell computergesteuerte Auswahlverfahren" ein (Eckhardt et al. 2016: 8). 47,5 Prozent der befragten Unternehmen gehen jedoch davon aus, dass „eine computergesteuerte und automatische Selektion von Bewerbungen in Zukunft immer häufiger zum Einsatz kommen wird" (a.a.O.). Dieselbe Studie kommt bei der Befragung von Stellensuchenden und Karriereinteressierten in Deutschland zu einem anderen Ergebnis: Vier von zehn Befragten gaben an, mindestens einmal im Rahmen der Stellensuche mit computergesteuerten und automatisierten Auswahlinstrumenten konfrontiert worden zu sein (a.a.O.). Das Aachener Unternehmen Precire analysiert 15 Minuten lange Sprachproben von Bewerbern mittels eines ADM-Prozesses, um auf Persönlichkeitsmerkmale zu schließen. Unternehmen wählen so Bewerber für Gespräche aus. Die Software soll auch für die Analyse von Stressbelastung im Unternehmen genutzt werden (Morrison 2017).

2.7 Überprüfbarkeit ermöglichen: Studienplatzvergabe in Frankreich

2009 führte das französische Bildungsministerium (Ministère de l'Education nationale, de l'Einseignement supérieur et de la Recherche) einen digitalen Zulassungsprozess für staatliche Hochschulen ein – Admission Post Bac (APB). Ein ADM-Prozess ordnet angehenden Studierenden seitdem einen Studienplatz und damit eine Hochschule zu. 2016 klagte die Schülervertretungsorganisation Droits des lycéens (Rechte der Gymnasiasten) auf Veröffentlichung des Algorithmus unter dem französischen Informationsfreiheitsgesetz (Commission d'accès aux documents administrati, CADA) (Thompson 2016).

2.7.1 Output: Zuordnung von Studienplätzen zu Abiturienten

2015 verteilte APB rund 740.000 Schülerinnen und Schüler auf mehr als 11.000 Studiengänge in ganz Frankreich. Im Ergebnis wurden 60 Prozent der Schüler ihrer erstgenannten Hochschule zugeordnet, 14 Prozent ihrer zweit- und 8 Prozent ihrer drittgenannten (Graveleau 2016). Ein Prozent der Studienplätze wurde im Losverfahren verteilt.

Die finale Entscheidung über die Zulassung oder Ablehnung eines Kandidaten liegt bei der Hochschule.

2.7.2 Datenbasis und Entscheidungslogik: Wohnort und Präferenz zählen

Jeder Schüler kann bis zu 24 Wunschstudiengänge angeben (jeweils verbunden mit einer bestimmten Hochschule), die er anschließend priorisieren muss. Auf dieser Basis berechnet der Algorithmus, für welche Studiengänge es mehr Kandidaten als Plätze gibt. Bei nicht ausgelasteten Studiengängen wird der Schüler seiner Studienwahl Nr. 1 zugeordnet. Bei überlasteten Studiengängen präferiert der Algorithmus zunächst diejenigen Studienbewerber, die ihren Schulabschluss im selben Schulbezirk (*académie*) abgelegt haben, in dem auch ihre Wunschhochschule liegt. In einem zweiten Schritt bewertet der Algorithmus die relative bzw. absolute Präferenz für einen Studiengang: Die absolute Präferenz entspricht der tatsächlichen Position des Studiengangs in der Wunschliste des Schülers, die relative der Position in der Wunschliste nach Abzug aller nicht ausgelasteten Studiengänge. Die Sortierung der Kandidaten priorisiert zunächst deren Wohnortnähe, dann die relative und schließlich die absolute Präferenz (Graveleau 2016) für einen Studiengang. Bleiben dann immer noch mehr Kandidaten übrig als Studienplätze zur Verfügung stehen, entscheidet ein Losverfahren. Alle Wunschlistenplätze unterhalb der höchstpriorisierten Zulassung werden automatisch ausgeschlossen und anderen Schülern angeboten („Further Education in France" o. J.).

Bekannt wurde diese Entscheidungslogik erst nach erheblichen Rechtsstreitigkeiten zwischen Droits des lycéens und dem französischen Bildungsministerium: So wurde der Code zwar schon im Januar 2016 angefordert, das Ministerium veröffentlichte ihn allerdings bewusst erst nach Verstreichen der 2016er-Fristen, um die Abiturienten „nicht zu beunruhigen" (Graveleau 2016). Im Juni 2016 wurde dann eine schematische Darstellung des Algorithmus veröffentlicht, die allerdings in keinster Weise überprüfbar war. Dies erklärte das Ministerium mit „Gründen der Sicherheit, insbesondere vor Hackerangriffen", aber auch mit dem Ziel des „leichteren Verständnisses", da der Quellcode mehr als 250 Seiten „unverständlichen Code" enthalte (a.a.O.). Erst nach mehreren Klagen veröffentlichte das Bildungsministerium schließlich im Oktober 2016 den Code des Algorithmus – der allerdings ohne Erläuterung der verwendeten Variablen wiederum nutzlos war und zudem ausgedruckt per Post geliefert wurde (Berne 2016). Droits des lycéens veröffentlichte die Dokumente dann auf der Kollaborationsplattform GitHub und bat Freiwillige um Unterstützung bei der Analyse des Programms auf. Ein Gesetzesentwurf, der nachträglich das Losverfahren sowie die verwendeten Kriterien juristisch unterfüttern sollte, wurde im Januar 2017 vom Ministerium überraschend wieder zurückgezogen (Stromboni 2017).

2.7.3 Konsequenzen: Implizite soziale Selektion mittels Wohnort und Wahlstrategie

Droits des lycéens kritisiert zwei Punkte: Erstens bedeutet die primäre Auswahl nach Wohnort, dass Schüler aus Paris höhere Chancen haben, an eine der renommierten *Grandes Écoles* zu kommen. Schüler aus ländlichen Regionen müssen dagegen erhebliche Nachteile in Kauf nehmen (Frouillou 2016; de Coustin 2016). Zweitens führen diese Kriterien sowie das Ausscheiden aller Optionen unterhalb der ersten Zulassung zu ausgefeilten Wahlstrategien: (Potenziell) überlastete Studiengänge werden möglichst hoch priorisiert, um nicht durch eine Zulassung in nicht überlasteten Studiengängen annulliert zu werden. Das wiederum suggeriert dem System einen „Ansturm

auf überlastete Fächer und führt dazu, dass im Losverfahren die Chance sinkt, einen der beschränkten Plätze zu bekommen, da zu viele Strohpuppen mitspielen (Graveleau 2016).

2.7.4 Evaluation: Hinweise auf Reproduktion sozialer Ungleichheiten und Ausweichstrategien

Die Ergebnisse der APB-Verteilung hat zum Beispiel Leila Frouillou (Sorbonne) untersucht (Lefauconnier 2016): Sie kommt zu dem Schluss, dass die Zuteilung der Studienplätze zwar allen gleichermaßen zugänglich sei, dadurch aber soziale Ungleichheiten reproduziert würden. Schüler aus den *académies* in Créteil und Versailles haben aufgrund des regionalen Zuteilungskriteriums wesentlich höhere Chancen auf eine Aufnahme in die vor allem in Paris beheimateten Eliteuniversitäten als Schüler aus Marseilles oder Lyon – ungeachtet ihrer Qualifikation. Da aber Wohnungen im Pariser Umfeld in der Regel wesentlich teurer sind als beispielsweise auf dem Land, wird somit implizit nach sozioökonomischem Hintergrund ausgewählt. Zudem entstanden durch das Bekanntwerden der Kriterien Ausweichstrategien, die nur wohlhabenden Studierenden offenstehen: So ziehen Eltern teilweise kurz vor dem Abitur ihres Kindes in den Einzugsbereich derjenigen *académie* um, die am ehesten in die gewünschte Universität führt. Andere wiederum nutzen die universitären Vorbereitungskurse (*prépas*), um die APB-Zuteilung zu vermeiden. Da diese jedoch ein bis zwei Jahre dauern, in dem beziehungsweise in denen der Schüler kein Einkommen hat, steht auch dieser Weg nur wohlhabenden Schülern offen (Frouillou 2016).

2.7.5 Zur Diskussion: Was nicht unabhängig überprüfbar ist, birgt Risiken systematischer Verzerrungen

Das Fallbeispiel zeigt den Wert der Überprüfbarkeit von Verfahren durch unabhängige Dritte. Denn wie bei der Studienplatzvergabe in Frankreich können auf den ersten Blick sinnvoll wirkende Kriterien in einem Auswahlverfahren zu systematischer Benachteiligung führen. Wenn der Wohnort der Eltern Einfluss auf die Zuteilung des Studienplatzes der Kinder hat, kann das soziale Ungleichheiten reproduzieren, wo Wohnort und sozialer Status korrelieren. Dieser Effekt ist auch bei anderen ADM-Prozessen zu beobachten. Zum Beispiel: Ein langer Arbeitsweg korreliert mit frühzeitiger Kündigung durch den Arbeitnehmer. Unternehmen könnten also Bewerbungen von Menschen mit weiter entfernt liegendem Wohnort ausschließen oder zurückstellen. Wenn aber der Wohnort mit dem sozialen Status korreliert, benachteiligt ein solches Verfahren systematisch (Trindel 2016).

Es ist unklar, ob der Zusammenhang zwischen Wohnort und sozialem Status bei dem Design des in Frankreich genutzten Zuteilungsverfahrens übersehen oder bewusst in Kauf genommen worden ist, weil in der Abwägung andere Gründe überwogen. In beiden Fällen ist die ursprüngliche Weigerung des Bildungsministeriums, den Code zu veröffentlichen, problematisch. Bei einem Versehen wurde so die Überprüfung durch Außenstehende erschwert. Sollte die Entscheidung in Abwägung bewusst getroffen worden sein, hätte dies vor Einführung des Zuteilungsverfahrens in einer breiten öffentlichen Debatte thematisiert werden müssen. Solche normativen Entscheidungen sollten nicht hinter der scheinbaren Objektivität von maschineller Entscheidungen versteckt werden.

Die Transparenz von ADM-Prozessen gegenüber der Öffentlichkeit kann Ausweichstrategien begünstigen. Das Bekanntwerden der Kriterien im vorliegenden Beispiel hatte ein verändertes Verhalten der Bewerteten zur Folge (z. B. den Ansturm auf potenziell überlastete Studiengänge ohne echtes Studieninteresse). Das könnte für Verfahren sprechen, die Erklärbarkeit und unabhängige Evaluation ohne komplette Veröffentlichung aller Verfahrensdetails ermöglichen. Das Argument des Bildungsministeriums hingegen, eine Veröffentlichung des Quellcodes würde Schwachstellen freilegen und somit Hacker einladen, scheint nur kurzfristig nachvollziehbar. Ausgebeutete Schwachstellen könnten gepatcht werden, der Algorithmus würde somit verbessert. Zudem hatte auch das Finanzministerium seinen Algorithmus zur Berechnung der Umsatzsteuer auf GitHub zur Verfügung gestellt (Berne 2016).

weitere Chancen	weitere Risiken
Ein gutes Verfahren könnte Abiturienten nachvollziehbar und transparent auf begrenzte (staatliche) Studienplätze verteilen.	Der Einfluss der Wohnorte auf die Verteilung führt in diesem Fall zu einer Diskriminierung über einen Stellvertreterwert. Vermögen und Einkommen der Eltern

	beeinflussen den Wohnort der Schüler, somit ist sozialer Status mittelbar ein Kriterium bei der Auswahl. Das Beispiel zeigt, dass bei der Gestaltung von ADM-Prozessen entscheidende Fehler bei der Messbarmachung auftreten können. Die Entscheidung, welche Daten erhoben und ausgewertet werden, ist nicht objektiv gegeben.
Das zentralisierte Vergabeverfahren entlastet Universitäten.	Das APB-Verfahren kann nicht umgangen werden, da staatliche Hochschulen ihre Studienplatzvergabe zentral darüber abwickeln. Hier zeigt sich ein grundlegendes Risiko von ADM-Prozessen: Die Entziehbarkeit sinkt. Da die Entscheidungslogik eines Systems mit vergleichsweise geringem Mehraufwand auf beliebig viele Fälle anwendbar ist, begünstigten solche Verfahren Zentralisierung. Dadurch skalieren aber auch Fehler stark.

2.7.6 Situation und Relevanz in Deutschland

Zwar wurde 2013 mit hochschulstart.de ein vergleichbares Portal auf den Weg gebracht – dank der sehr individualisierten, dezentralen Entscheidungskriterien der Hochschulen, welche Schüler sie annehmen, ist die Gefahr eines dominanten Einheitsprozesses aber gering. Auch die Bedeutung der Wohnortnähe dürfte in Deutschland aufgrund der fehlenden lokalen Konzentration von Spitzenunis geringer ausfallen.

Im Fach Medizin erstellte 2015 das Karlsruher Institut für Technologie (KIT) ein sogenanntes „Zulassungsorakel", das für jeden Studierenden dessen individuelle Wartezeit auf einen Medizinstudienplatz errechnet. Dazu nutzten die Wissenschaftler öffentlich zugängliche Daten von hochschulstart.de selbst: unter anderem Abiturnote und -jahrgang, die Anzahl der Wartesemester, ob und mit welcher Note die Abiturienten den Medizinertest abgelegt haben sowie ob in der Zwischenzeit eine (berufsrelevante) Ausbildung absolviert wurde. Das Forscherteam verglich rund 250.000 Studierendenprofile sowie die Zulassungskriterien an 36 Hochschulen mit der jeweiligen Zulassungsentscheidung. Ein maschinell lernender Algorithmus errechnete aus diesen Daten dann die für die Zulassung relevanten Gewichtungen (Jung et al. 2015). Ziel der Wissenschaftler war es, die Unsicherheit der Studierenden rund um ihre Bewerbung auf einen Studienplatz in Medizin zu minimieren und so unnötige Wartezeiten auf das oder eine mögliche Abwendung vom Medizinstudium zu verhindern.

Ein vergleichbarer Versuch, staatlich verwendete ADM-Prozesse per Gerichtsverfahren zu befreien, wurde bisher nicht dokumentiert.

2.8 Soziale Wechselwirkungen beachten: Ortsbezogene Kriminalitätsprognosen

Auf ADM-Prozessen beruhende Prognosen für die Polizeiarbeit können den Fokus von Ermittlern auf Personen lenken (siehe Kapitel 2.2). Ein anderer Ansatz konzentriert sich auf Orte, die ADM-Prozesse prognostizieren – beispielsweise Hotspots für bestimmte Delikte wie Wohnungseinbrüche. Zu den bekannten kommerziell verfügbaren Analyseprogrammen mit diesem Ansatz zählen Precobs (Institut für musterbasierte Prognosetechnik, Deutschland) und Predpol (Predpol, USA).

Mit geographischen Mustern von Kriminalität arbeiten Behörden und Kriminologen seit dem 19. Jahrhundert, damals zuerst in London. Lange bevor in den nuller Jahren Software zum Einsatz kam, erarbeiteten menschliche Analysten aus Kriminalitätsstatistiken ortsbezogene Kriminalitätsschwerpunkte: „For example, half of the crime in Seattle over a fourteen-year period could be isolated to only 4.5% of city streets. Similarly, researchers in Minneapolis, Minnesota found that 3.3% of street addresses and intersections in Minneapolis generated 50.4% of all dispatched police calls for service“ (Gluba 2014: 5). Solche rückblickenden Analysen sind eine Grundlage erkenntnisbasierter Polizeiarbeit.

2.8.1 Output: Einbruchsprognosen für 250 x 250 Meter große Flächen

Die Precobs-Software visualisiert die Prognose für Einbrüche am Tag in 250 mal 250 Meter großen Zielgebieten (Brühl und Fuchs 2014). Das System veranschaulicht die Prognosen durch eine farbliche Markierung der Kacheln: In roten Kacheln ist die erwartete Folgedeliktquote am höchsten, in gelb, grün und blau eingefärbten graduell niedriger. Ein „near repeat“ liegt für Precobs vor, wenn „mindestens zwei Delikte aus einem Deliktfeld innerhalb von 72 Stunden in einem eingegrenzten geografischen Raum auftreten“ (Institut für musterbasierte Prognosetechnik 2014).

Die in den Vereinigten Staaten und Großbritannien verbreitete Software Predpol liefert Prognosen in einem ähnlichen Format: Das Gebiet wird in 150 mal 150 Meter große Quadrate eingeteilt, für jedes davon errechnet Predpol das Deliktrisiko für die folgenden zwei Schichten von jeweils zwölf Stunden und zeigt dann nur die 20 Quadrate mit dem höchsten Risiko auf der Karte (Mohler et al. 2015: 10).

Letztlich entscheiden die Polizisten, ob und wie sie Ressourcen in einer Schicht auf Basis der Hotspot-Prognosen verteilen.

2.8.2 Datenbasis und Entscheidungslogik: Nicht personenbezogene Tateigenschaften

Die Prognose von Hotspots für bestimmte Delikte basiert auf der Near-Repeat-Theorie. Dieser kriminologische Ansatz geht davon aus, dass bei Delikten wie Autodiebstahl, Wohnungs- und Autoeinbrüchen die Wahrscheinlichkeit für weitere Taten im örtlichen Umfeld nach einem Delikt steigt. Für Wohnungseinbrüche zeigen empirische Studien aus Großbritannien, den USA, den Niederlanden, Neuseeland und Australien statistisch signifikante Near-Repeat-Muster (Ferguson 2012: 19).

Precobs analysiert wenige Parameter wie Tatzeit, Beutetyp (z. B. Bargeld, Geldkassette), Gebäudetyp (z. B. Büro-, Geschäfts-, Wohngebäude) oder Vorgehen beim Einbruch (z. B. hebeln, drücken, treten), um Muster von Serientätern zu erkennen (Brühl 2014). Die Daten sind nicht personenbezogen, Angaben zu Tätern oder Opfern fließen nicht in die Analyse ein (Schindler und Wiedmann-Schmidt 2015). Trotz dieses klar begrenzten Ziels und Inputs sagt Ralf Middendorf, einer der Precobs-Entwickler: „Wir sehen Zusammenhänge, die wir nicht erklären können“ (Brühl 2014).

2.8.3 Konsequenzen: Größere Polizeipräsenz im Risikogebiet

Polizeibehörden nutzen Precobs und Predpol, um Prioritäten beim Einsatz von Streifen zu setzen. Zum Beispiel im britischen Kent und in Los Angeles (Mohler et al. 2015: 10) oder beim Precobs-Pilotprojekt in München: „Berechnet Precobs, dass an einem Tag in einer Region mit Einbrüchen zu rechnen ist, verstärkt die Polizei dort ihre Streifenpräsenz“ (Brühl und Fuchs 2014).

2.8.4 Evaluation: Gemischte Ergebnisse zur Wirksamkeit

Die wenigen in Peer-Review-Verfahren überprüften Untersuchungen zum Einsatz ortbezogener Kriminalitätsprognosen haben sehr unterschiedliche Ergebnisse. Eine von Mitarbeitern des Unternehmens Predpol verfasste Studie zum Predpol-Einsatz in Kent (Großbritannien) und Los Angeles kommt zu dem Ergebnis, dass die Software bis zu zweimal so viel Einbrüche korrekt verortet vorhersagte wie menschliche Analysten: „Our results show that ETAS models predict 1.4-2.2 times as much crime compared to a dedicated crime analyst using existing criminal intelligence and hotspot mapping practice“ (Mohler et al. 2015:1402). Zudem sank in Folge die Kriminalitätsrate an den von der Predpol-Software prognostizierten Hotspots, nachdem dort zusätzliche Kräfte auf Streife geschickt wurden.

Als eindeutiger Beleg für die Wirksamkeit von Predpol taugen diese Ergebnisse allerdings nicht, da die Qualität der menschlichen Entscheidungen von nur vier Analysten in der Vergleichsgruppe unbekannt ist und andere Nebenwirkungen denkbar sind:

> *„But with only four human analysts of unknown effectiveness included in the study, the comparison is not wholly convincing. (…) More patrol time on ETAS hot spots could indeed be reducing crime; then again, on days when there is little crime for whatever reason, officers could have more time to visit suspect areas“ (Perkowitz 2016).*

Im Jahr 2011 ermittelte die wissenschaftliche Evaluation in einem anderen Pilotprojekt mit anderer Software keine Vorzüge der ADM-Lösung: „This study found no statistical evidence that crime was reduced more in the experimental districts than in the control districts“ (Hunt, Saunders und Hollywood 2014).

In Mailand nutzt die Polizei ein ADM-Verfahren zur Prognose von Hotspots für Diebstähle in einigen Stadtteilen. Die Gendarmerie, die für andere Stadtteile zuständig ist, nutzt solche Verfahren nicht. Der Vergleich der Aufklärungsquote in den Vierteln soll zeigen, dass sich seit dem Einsatz der Prognosesoftware die Aufklärungsquote im von der Polizei betreuten Gebiet um acht Prozentpunkte verbessert hat (Mastrobuoni 2014: 7) – die Studie ist allerding unveröffentlicht.

Es ist unklar, ob die Kriminalität infolge verstärkter Streifen an Hotspots tatsächlich sinkt oder abwandert. „Fragen der schlichten Verschiebung von Kriminalität durch die Konzentration auf bestimmte Areale, wie sie Predictive Policing beinhaltet, sind aufgrund der Forschungslage nicht hinreichend beantwortet“ (Gluba 2014: 11).

2.8.5 Zur Diskussion: Soziale Wechselwirkung erschwert Folgenabschätzung

Zutreffende Vorhersagen der Hotspots von Einbrüchen wie im Fallbeispiel können bei verstärkter Polizeipräsenz dort bestimmte Delikte sogar verhindern. Die Risiken sind kleiner als bei personenbezogenen Prognosen (Selbst 2016: 21). Allerdings zeigen bei diesem Fallbeispiel die vorliegenden Analysen, wie komplex die Wechselwirkungen eines ADM-Verfahrens mit der Umwelt sein können. Schon bei einem derart eingeschränkten Einsatzgebiet wie der ortsbasierten Prognose weniger Deliktarten wird deutlich: Nur die Analyse der Wirkungen des gesamten sozioinformatischen Prozesses kann zeigen, in welchem Verhältnis Chancen und Risiken stehen. Zu klären sind Fragen wie:

- Verdrängt ortsbezogenes „predictive policing“ Kriminalität an andere Orte?
- Entstehen systematische Verzerrungen, weil die Prognosen auf Statistiken zu der Polizei bereits bekannten Fällen beruhen und das Verfahren daher die Aufmerksamkeit eher auf das Hell- als auf das Dunkelfeld lenkt?
- Passen Täter ihr Vorgehen den genutzten Verfahren an?

weitere Chancen	weitere Risiken
Auch menschliche Analysten orientieren sich an ortsbezogenen Einschätzungen, Statistiken und Hinweisen von Informanten. Die Systematisierung und Evaluation solcher Quellen für den Einsatz in ADM-	Nur bestimmte Delikte sind territorial. Es besteht das Risiko, dass mehr Ressourcen für die Verfolgung dieser prognostizierbaren Delikttypen investiert werden, weil polizeiliche Erfolge schneller möglich sind (Selbst 2016: 17).

Prozessen kann zu einer öffentlichen Diskussion solcher bisher nicht sichtbarer Entscheidungen führen (Prabhu 2015: 11).

2.8.6 Situation und Relevanz in Deutschland

Ortsbasierte Kriminalitätsprognosen sind in Europa im regulären Einsatz, zum Beispiel in Zürich (Baumgartner, 2015), Kent (Mohler et al. 2015) und Mailand (Mastrobuoni 2014). In Deutschland ist diese Form des Predictive Policing derzeit in 14 Pilotprojekten und Tests im Einsatz oder in der Entwicklung (Pilpul 2016).

2.9 Zweckentfremdung verhindern: Kreditscoring in den USA

In den Vereinigten Staaten bestimmen drei große nationale Auskunfteien den Markt für Prognosen der Kreditwürdigkeit von Privatkunden: TransUnion, Experian und Equifax. Jeder dieser Anbieter verarbeitet laut der US-Verbraucherschutzbehörde Consumer Financial Protection Bureau (CFPB) monatlich etwa 1,3 Milliarden Aktualisierungen der Profile von mehr als 200 Million Verbrauchern (Hurley und Adebayo 2016: 154). Diese Auskunfteien bieten Prognosen basierend auf traditionellen Modellen wie dem sogenannten Fico-Score an, die ausschließlich die Kredithistorie und relevante Gerichtsurteile beispielsweise zu Insolvenz oder Zwangsvollstreckung als Datengrundlage einbeziehen. Diese Prognosemodelle schließen nach Schätzungen des Interessenverbandes National Credit Reporting Association ungefähr 70 Millionen US-Bürger von Krediten aus, die mangels Daten keinen Score erhalten oder allein aufgrund beschränkter Informationen eine schlechte Rückzahlprognose erhalten (Robinson und Yu 2014: 6). Zu den „unscoreables“ (Menschen mit zu wenig Daten für traditionelle Modelle) können laut Einschätzung von Experian zum Beispiel Einwanderer oder College-Absolventen zählen (Hurley und Adebayo 2016: 155).

2.9.1 Output: Personenbezug herausgerechnet, Transparenzpflicht verloren

Traditionelle Prognosemodelle der Anbieter Fico und Vantagescore geben eine Zahl aus. Je höher der Wert, desto höher wird die Kreditwürdigkeit eingeschätzt. Zwar interpretieren die Auskunfteien selbst die Werte für ihre Bewertungen von Kreditnehmern. Der ADM-Anbieter Vantagescore gibt dennoch öffentlich an, dass ein Kreditnehmer in der besten Stufe („prime“) einen Score zwischen 661 und 780 erzielt (Vantagescore 2013).

Der Output beeinflusst auch, ob Prognosen unter die Transparenzanforderung der FCRA-Regulierung zu Kreditauskünften fallen. Da das Gesetz auf Individuen abzielt, erfasst es nicht aggregierte Marketing-Scores, die beispielsweise Aussagen über Häuserzeilen machen.

> *„Aggregated marketing scores – which are computed on a household or block level, and arguably not tied to any one consumer's identity – have become a primary way for credit bureaus to sell, and for creditors and other actors to use, consumers' credit histories to market to them with greater precision. These products often come within a hair's breadth of identifying a person. (…) In other words, it provides detailed insight into the financial characteristics of the ‚group‘ of people in a single household – and does so putatively without triggering any of the protections of the FCRA“ (Robinson und Yu 2014: 17).*

Potenzielle Kreditgeber können in ihre Entscheidung auch andere Faktoren einbeziehen. Wie Fico in einer Kundeninformation angibt: „Your credit score is calculated from your credit report. However, lenders look at many things when making acredit decision such as your income, how long you have worked at your present job and the kind of credit you are requesting” (Fair Isaac Corporation 2017). Es ist unklar, wie oft solche Entscheidungen automatisiert ablaufen und wann ein menschlicher Entscheider einbezogen wird. Bei einigen Anwendungen ist offenkundig, dass die Urteile rein maschinell und automatisiert fallen müssen: Beispielsweise bei Online-Kreditanträgen oder der Bewertung der wartenden Anrufer in einer Hotline auf Basis ihres Kredit-Scorings.

2.9.2 Datenbasis und Entscheidungslogik: Kreditgeber, Stromversorger, soziale Netzwerke

Zwei Gesetze definieren, welche Prognosen zulässig sind: Der Fair Credit Reporting Act (FCRA) schreibt vor, dass die verkauften Daten über Individuen relevant und zutreffend sein müssen und nur zu bestimmten erlaubten Zwecken verwendet werden dürfen. Der Equal Credit Opportunity Act (ECOA) verbietet es, in Systemen zur Bewertung der Kreditwürdigkeit geschützte Eigenschaften wie Rasse oder Alter einzubeziehen (Robinson und Yu 2014b: 6).

Zwei grundlegende Veränderungen prägen die Situation in den USA:

- Neue Prognosemodelle nutzen Datenquellen wie soziale Netzwerke oder Konsumprofile, auch um die Bonität bisher nicht bewerteter Menschen vorherzusagen (siehe: Datenbasis und Entscheidungslogik:).
- Auskunfteien entwickeln neue Prognosen, die konterintuitiv nicht unter die FCRA-Regulierung fallen: So können sogenannte Marketing-Scores auf Informationen über die Kreditwürdigkeit beruhen, werden aber nicht zur Kreditvergabe, sondern zum Beispiel zur Preisgestaltung genutzt (a.a.O.).

Robinson und Yu (2014b: 4) unterscheiden drei Ansätze zur Bonitätsprognose in den USA anhand der in den ADM-Prozessen genutzten Daten:

- Traditionelle Prognosemodelle verwenden nur Informationen über die Rückzahlung von Krediten, beispielsweise von Kreditkartenfirmen oder Immobilienkreditgebern.
- Etablierte alternative Modelle („mainstream alternative models“) werten auch Daten aus, welche von Mitgliedsfirmen der Auskunfteien stammen, sich aber auch auf die regelmäßige Begleichung von Forderungen beziehen, zum Beispiel bei Versorgern (Strom, Wasser usw.) (a.a.O.: 10).
- Neuartige alternative Modelle („fringe alternative models“) nutzen auch Daten für Bonitätsprognosen, die sich überhaupt nicht auf die Begleichung von Forderungen beziehen. Je nach Modell können das beispielswiese Social-Media-Profile sein, Positionsdaten des Smartphones des Anfragenden, Informationen über das Einkaufsverhalten oder Auswertungen, wie schnell ein Nutzer Informationen aus der Website des Kreditgebers durchscrollt (a.a.O.: 13 ff.). Diese Modelle werden in den US oft auch als „alternative credit decisioning tools“ (ACDT) bezeichnet.

Die Entscheidungslogik der Systeme ist schon bei traditionellen Prognosemodellen schwer nachvollziehbar. Die beiden großen Anbieter traditioneller Prognosemodelle Fico und Vantagescore betreiben viele unterschiedliche Versionen ihrer ADM-Prozesse für verschiedene Kunden (Hurley und Adebayo 2016: 155). Die Informationen beider Anbieter über die Funktionsweise differenzieren nicht zwischen unterschiedlichen Verfahren:

- Fico gibt an, dass der Scoringwert auf Informationen zum individuellen Verhalten auf fünf Kategorien beruht: bisherige Zahlungshistorie, Umfang der Kredithistorie, Höhe bestehender Kredite, Art der bestehenden Kredite, Menge neuer Kreditanfragen. Die Darstellung beziffert das Gewicht der Kategorien zwar mit Prozentangaben, relativiert diese Aussagen aber. Die Gewichtung sei bei jedem Scoring eine andere: „… it’s impossible to measure the exact impact of a single factor in how your credit score is calculated without looking at your entire report. Even the levels of importance shown in the FICO Scores chart are for the general population, and will be different for different credit profiles“ (Fair Isaac Corporation 2017).
- Vantagescore gibt dieselben Grundlagen der Bonitätsprognose an wie Fico und ergänzt zudem noch ausgeschöpfte Kreditlinien. Kreditnehmer erhalten Verhaltenstipps für ein gutes Scoring. Zum Beispiel: „Maintain a mix of accounts (credit cards, auto, mortgage) over time to improve your score. Prime consumers have an average of 13 loans. Typically the oldest loan is more than 15 years old“ (Vantagescore 2013) Das ist ein möglicher Hinweis auf die Entscheidungslogik: Der ADM-Prozess könnte die Ähnlichkeit des Profils eines Kreditsuchenden mit den Profilen verlässlich zurückzahlender Kreditnehmer vergleichen.

Ein wichtiger Faktor bei allen etablierten Modellen ist das Alter der vorliegenden Daten: „Credit files that have gone more than six months with no reported activity are considered ‚stale‘ by the FICO algorithm, and will not produce a score“ (Robinson und Yu 2014c: 17). Dieser Umstand kann teilweise erklären, warum ungefähr 70 Millionen US-Bürger mangels Daten keine Bonitätsprognose erhalten.

Citron und Pasquale kritisieren, dass diese Informationen die Entscheidungslogik der Verfahren nicht nachvollziehbar machen und für Betroffene im Einzelfall wenig hilfreich sind: „Looking forward, a consumer has no idea, for example, whether paying off a debt that is sixty days past due will raise her score. The industry remains highly opaque, with scored individuals unable to determine the exact consequences of their decisions“ (2014: 18).

Die Entscheidungslogiken alternativer Prognosemodelle sind auf Grundlage öffentlich zugänglicher Informationen noch schwerer erklärbar. ZestFinance, ein Anbieter neuartiger Prognosemodelle, verarbeitet bis zu 10.000 Datenpunkte je Kreditantragsteller, darunter Mobilfunkzahlungen, aber auch Verhaltensdaten wie „unusual observations, such as whether applicants use proper spelling and capitalization on their application form, how long it takes them to read it, and whether they bother to look at the terms and conditions“ (O’Neil 2016a: 144). Der Gründer und

Geschäftsführer des Anbieter Douglas Merrill erweckt in einem Interview den Eindruck, dass die Entscheidungslogik auch für sein Unternehmen nicht in jedem Einzelfall nachvollziehbar ist: „Merrill acknowledges that in many cases, there's no explanation for why a particular data point helps or hurts a credit score. For instance, borrowers who write in all-caps are riskier, the firm's credit scoring system discovered after underwriting thousands of loans. ‚We don't know why. It just is', said Merrill" (Koren 2015).

2.9.3 Konsequenzen: Scoring beeinflusst Versicherungsprämien und Bewerberauswahl

Viele US-Bürger erhalten keine oder nur sehr teure Kredite, weil ihr Profil nicht datenreich genug für Prognosen ist.

Das kann umso schwerer wiegen, da die ursprünglich für eine Prognose der Ausfallwahrscheinlichkeit von Krediten entwickelten Verfahren auch als indirekte Anzeiger bei ganz anderen Fragen genutzt werden. Einige Beispiele:

Versicherungen: Bonitätsprognosen haben in vielen US-Bundesstaaten einen Einfluss auf die Kosten von Autoversicherungen. Ein unterdurchschnittlicher Scoringwert kann in bestimmten Fällen die Prämien um bis zu 1301 Dollar im Jahr verteuern – unabhängig vom Fahrverhalten, ergab eine Preisanalyse der Verbraucherorganisation Consumer Reports (Consumer Reports 2015). Die Praxis ist in allen US-Bundesstaaten außer Kalifornien, Hawaii und Massachusetts zulässig. In einigen Staaten können die Preisaufschläge für schlechte Bonitätsprognosen höher ausfallen als für Verurteilungen wegen Alkohols am Steuer (O'Neil 2016a: 149).

Bewerberauswahl: 47 Prozent der Personalabteilungen nutzten laut einer 2012 durchgeführten Umfrage des US-Brachenverbandes Society for Human Resource Management (SHRM) Bonitätsprognosen bei der Bewerberauswahl: 34 Prozent der Befragten für einige, 13 Prozent für alle Kandidaten (Society for Human Resource Management 2012: 8). Betroffen sind laut einer US-Hilfsorganisation vor allem Jobs im Niedriglohnsektor: „The people contacting her group, she said, are ‚mostly lower-wage workers", especially those applying to big retail chains" (Rivlin 2013).

Arbeitslosigkeit ist die Folge für Betroffene:

> *„Among survey respondents who are unemployed, 1 in 4 says that a potential employer has requested to check their credit report as part of a job application. 1 in 10 survey respondents who are unemployed have been informed that they would not be hired for a job because of the information in their credit report. Among job applicants with blemished credit histories, 1 in 7 has been advised that they were not being hired because of their credit" (Traub 2013: 9).*

2.9.4 Evaluation: Kaum unabhängige Studien, Hinweise auf Altersdiskriminierung

Eine aktuelle, unabhängige, repräsentative und systematische Untersuchung der Qualität der unterschiedlichen Prognosemodelle existiert nicht. Die 2012 von der US-Zentralbank veröffentlichte Untersuchung zu möglichen ungleichen Auswirkungen der Scoringwerte basiert auf Datensätzen aus den Jahren 2003 und 2004 (Avery, Brevoort und Canner 2012: 3). Ergebnis dieser 300.000 mit demographischen Informationen angereicherten Datensätzen: Es ist keine Ungleichbehandlung nach Ethnie oder Geschlecht erkennbar, aber es gibt Hinweise darauf, dass die Prozesse junge Menschen benachteiligen:

> *„Our results provide little or no evidence that the credit characteristics used in credit history scoring models operate as proxies for race or ethnicity. (…) We do, however, find some evidence that credit characteristics associated with the length of an individual's credit history (…) may have a disparate impact by age. In particular, we find that the predictiveness of this credit characteristic increases when the credit scoring model is estimated in an age neutral environment" (Avery, Brevoort und Canner 2012: 27).*

Dieses Ergebnis zeigt eine Schwäche der traditionellen Prognosemodelle auf: Wer wenig hat (Lebensjahre, Daten, Zahlungshistorie), dem werden teurere Kredite gegeben. Oder gar keine, wenn zu wenig Daten vorliegen.

Die Zuverlässigkeit der traditionellen Prognosemodelle hat 2012 die US-Verbraucherschutzbehörde FTC untersucht: 1001 Studienteilnehmer bewerteten ihre Kreditauskünfte (2968 insgesamt). Ergebnisse: 26 Prozent der Teilnehmer entdeckten Fehler in ihren Auskünften, 21 Prozent haben durch Anfechtung dieser Informationen eine

Korrektur erreicht, nur bei 13 Prozent änderte sich nach dieser Korrektur der Scoringwert (Federal Trade Commission et al. 2012: 5).

Wie das Einbeziehen von Zahlungen an Strom- und Telekommunikationsversorger auf Bonitätsprognosen wirkt, untersuchte das Policy and Economic Research Council (PERC). Das Urteil über etablierte alternative Modelle fällt positiv aus. So wurden 25 Prozent der untersuchten Menschen, die zuvor zu wenig Zahlungsinformationen für traditionelle Prognosemodelle hatten („thin-file population") nach dem Einbeziehen der Versorgerdaten in eine bessere Risikokategorie eingestuft worden. Nur sechs Prozent hätten nach der erweiterten Auswertung eine schlechtere Einstufung in eine niedrigere Risikostufe erhalten (Turner et al. 2012: 6).

Die Anbieter neuartiger alternativer Prognosemodelle versprechen, weit mehr Menschen Zugang zu Krediten zu ermöglichen. Zu den dabei genutzten Modellen liegen allerdings keine unabhängigen, repräsentativen und systematischen Evaluationen vor, wie Robinson und Yu (2014) bilanzieren:

> *„Less still is known about the financial startup scene, which relies on even more exotic data. For example, ZestFinance boasts that its "big data underwriting model provides a 40% improvement over the current best-in-class industry score." But it is unclear how accurate the "best-in-class industry score" actually is for Zest's target population of consumers, much less how ZestFinance measures up to that benchmark" (a.a.O.: 27).*

2.9.5 Zur Diskussion: Zweckentfremdung von Scoringwerten überträgt Teilhabeeffekte

Das Fallbeispiel zeigt, welche Auswirkungen es hat, wenn ein Konzept wie Kreditwürdigkeit als Proxy-Wert in vielen anderen Lebensbereichen zur Klassifizierung von Menschen dient. In einigen der Fälle wird das Konzept Kreditwürdigkeit klar zweckentfremdet, zum Beispiel, um Menschen mit schlechtem Scoring teure Kreditkarten anzubieten oder ihre Anrufe in Call-Centern hintenanzustellen (O'Neil 2016b: 132 ff.).

Die automatisierte Abwicklung von ADM-Verfahren erleichtert solche Beispiele von Zweckentfremdung. Der gesetzliche Rahmen in den Vereinigten Staaten hat dazu geführt, dass Versicherungen automatisierte Verfahren personenbezogene Bonitätsauskünfte in sogenannten Marketing Scores umwandeln können. Diese fallen in den USA nach gegenwärtig dominierender Auffassung nicht unter die Kreditregulierung, aber reichen für eine letztlich personalisierte Ungleichbehandlung durch Unternehmen aus (Robinson und Yu 2014b: 6).

Der Scoringwert der Kreditwürdigkeit liefert eine eindeutige, leichte Antwort – aber in den Beispielen für Zweckentfremdung nicht auf die eigentlich relevante Frage. Ob die Kreditwürdigkeit einer Person wirklich etwas über die Arbeitsleistung oder das Unfallrisiko aussagt, ist mehr als zweifelhaft. So werden Nachteile aus einem Lebensbereich in andere übertragen. Diese Beispiele zeigen, dass sich die Auswirkungen von ADM-Prozessen auf Individuen nicht allein aus einer Datenschutzlogik heraus bewertet lassen.

weitere Chancen	weitere Risiken
ADM-basierte alternative Modellen für Bonitätsprognosen können Menschen einen Zugang zu Krediten eröffnen, die bisher mangels Informationen über ihr Zahlungsverhalten von traditionellen Modellen gar nicht oder als hohes Risiko bewertet werden (Turner et al. 2012: 23; Hurley und Adebayo 2016: 156).	Die alternativen Modelle sind kaum unabhängig erforscht (Robinson und Yu 2014: 27). Sie könnten auch einige Bevölkerungsgruppen systematisch diskriminieren, zum Beispiel Menschen mit Rechtschreibschwächen, wenn Orthographiefehler im Kreditantrag als Signal für ein erhöhtes Ausfallrisiko gewertet werden (O'Neil 2016a: 144).

2.9.6 Situation und Relevanz in Deutschland

In Deutschland bieten Auskunfteien Informationen über Unternehmen und Personen an. Das Bundesdatenschutzgesetz erlaubt ein Scoring unter bestimmten Voraussetzungen. Zum Beispiel müssen die zur Berechnung des Wahrscheinlichkeitswerts genutzten Daten unter Zugrundelegung eines wissenschaftlich anerkannten mathematisch-statistischen Verfahrens nachweisbar für die Berechnung der Wahrscheinlichkeit des bestimmten Verhaltens

erheblich sein, es dürfen nicht ausschließlich Anschriftendaten genutzt werden (§ 28b Bundesdatenschutzgesetz, BDSG).

Der Bundesgerichtshof (BGH) hat 2014 entschieden, dass bewertete Menschen keinen Anspruch haben zu erfahren, wie die Bewertung ihres zukünftigen Verhaltens genau berechnet worden ist. In dem Verfahren klagte eine Frau, der nach einem Scoring durch die Schutzvereinigung für allgemeine Kreditsicherung (Schufa) kein Kredit gewährt wurde, auf Auskunft. Ihrer Ansicht nach genügte die von der Schufa übermittelte Datenübersicht nicht den gesetzlichen Anforderungen. Der BGH urteilte, dass die „abstrakte Methode der Scorewertberechnung nicht mitzuteilen ist“ und dass als Geschäftsgeheimnis unter anderem diese Informationen geschützt sind: „(…) die im ersten Schritt in die Scoreformel eingeflossenen allgemeinen Rechengrößen, wie etwa die herangezogenen statistischen Werte, die Gewichtung einzelner Berechnungselemente bei der Ermittlung des Wahrscheinlichkeitswerts und die Bildung etwaiger Vergleichsgruppen als Grundlage der Scorekarten“ (Bundesgerichtshof 2014: 1). Gegen dieses Urteil ist eine Verfassungsbeschwerde anhängig („Schufa-Klägerin zieht vor Verfassungsgericht“ 2014).

3 Fazit

Die Fallbeispiele zeigen, dass und wie ADM-Prozesse Entscheidungen über Menschen beeinflussen. Wenn Maschinen uns bewerten und ihre Prognosen – wie beim Einsatz vor Gericht oder durch die Polizei – Freiheitsrechte oder – wie bei der Bewerberauswahl oder bei Bonitätsprognosen – Gleichstellung berühren, muss die Gesellschaft über Fairness und Teilhabewirkung dieser Verfahren diskutieren.

Hier lohnt sich der genaue Blick auf den konkreten Kontext, denn nicht jeder ADM-Prozess ist gleich riskant: Die gesellschaftlichen Anforderungen an ADM-Prozesse können variieren, abhängig von der Wirkung dieser Verfahren auf die Gesellschaft und die individuellen Grundrechte. Rechtschreibkorrektur oder Navigationssysteme haben andere Konsequenzen auf das Leben eines Menschen als Systeme, die ihm Kredit- oder Delinquenzrisiken zuschreiben.

Die Aggregation der in den Fallbeispielen aufscheinenden Chancen und Risiken (siehe Tabelle 2: Abstrahierender Überblick von Chancen und Risiken aus den Fallbeispielen auf der Folgeseite) weist auf einige übergreifende teilhabekritische Faktoren von ADM-Prozessen hin. Diese liegen auf unterschiedlichen Aspekten des gesamten sozioinformatischen Prozesses und betreffen unterschiedliche Ebenen. Drei Beispiele:

- *Gestaltung von ADM-Prozessen auf Mikro- und Mesoebene:* Schon die Auswahl von Daten und die Bestimmung von Kriterien zu Beginn des Entwicklungsprozesses können normative Setzungen enthalten, die in bestimmten Fällen grundsätzliche gesellschaftliche Fragen berühren.

- *Anbieter- und Betreiberstruktur auf Makroebene:* Die Vielfalt unterschiedlicher ADM-Prozesse und Betreiber kann Teilhabe stärken (z. B. durch Bonitätsprognosen, die sich an Menschen richten, die bisher aus dem System gefallen sind) sowie Entzugsmöglichkeiten und Falsifizierungspotenzial vergrößern. Entsprechend erhöhen monopolistische Strukturen das Risiko für den Einzelnen, „aus dem System zu fallen“.

- *Umgang mit ADM-Prognosen auf Mikro-, Meso- und Makroebene:* Das Zusammenspiel von Technik, Gesellschaft und Individuen hat großen Einfluss auf den Umgang mit und somit die Wirkung von Algorithmen. Wichtige Fragen sind daher: Wie gehen Menschen (sowohl ADM-Entwickler als auch -Anwender und die Bevölkerung) mit den automatisierten Prognosen um? Sind in diesen Verfahren Möglichkeiten angelegt, die ADM-Vorhersagen zu widerlegen?

Nötig sind hier weitere systematische Analysen möglicher Fehlerquellen auf unterschiedlichen Stufen von ADM-Prozessen – von der Definition der Ziele, über die Messbarmachung der Konzepte, die Datenerhebung und Algorithmusauswahl bis hin zur Einbettung des Verfahrens den gesellschaftlichen Kontext (vgl. Zweig 2016). Es braucht Gütekriterien für ADM-Prozesse, die alle Ebenen und Stufen einbeziehen. Die in der Einleitung hervorgehobenen Handlungsbedarfe können hierfür eine erste Grundlage bieten (vgl. Tabelle 1: Fehlerquellen in Prozessen algorithmischer Entscheidungsfindung).

Ein wichtiges geradezu konstitutives Qualitätsmerkmal soll an dieser Stelle noch einmal besonders betont werden: Die Analyse der Chancen, Risiken und gesellschaftlichen Folgen war nur möglich, weil unabhängige Dritte die Güte der maschinellen Entscheidungen prüfen konnten. Institutionen wie das Recherchebüro Propublica (siehe Kapitel 2.1 Rückfallprognosen vor Gericht), der US-Rechnungshof (siehe Kapitel 2.5 automatische Gesichtserkennung) oder die Schülervertretungsorganisation Droits des lycéens (siehe Kapitel 2.7 Studienplatzvergabe in Frankreich) haben Zeit und Geld in Datenbeschaffung, Datenauswertung und juristische Auseinandersetzung investiert, um Erklärbarkeit und Nachvollziehbarkeit der jeweils eingesetzten Algorithmen herzustellen. Von solchen Institutionen hängt derzeit ab, ob eine gesellschaftliche Debatte über die Wirkung bestimmter ADM-Prozesse überhaupt möglich ist. Das muss sich ändern. Denn Überprüfbarkeit und Nachvollziehbarkeit algorithmischer Entscheidungen sind eine unverzichtbare Erkenntnisgrundlage für einen lösungsorientierten gesellschaftlichen Diskurs mit dem Ziel, dass ADM-Prozesse für mehr Teilhabe gestaltet werden und maschinelle Entscheidungen den Menschen dienen.

Tabelle 2: Abstrahierender Überblick von Chancen und Risiken aus den Fallbeispielen (Quelle: eigene Darstellung)

Wirkungsdimension	Chancen	Risiken
Normative Setzungen	Bei der Gestaltung eines ADM-Prozesses müssen normative Entscheidungen (z. B. über Fairness-Kriterien) vor dem Einsatz fallen. Das bietet die Möglichkeit, früh und grundlegend ethische Fragen öffentlich zu diskutieren und Entscheidungen zu dokumentieren.	ADM-Prozesse können normative Entscheidungen im Design verstecken. Wenn erst nach Abschluss der Gestaltung Raum für Diskurs ist, werden Setzungen leichter als gegeben akzeptiert.
Datenbasis	Software kann eine deutlich größere Datenbasis analysieren als Menschen und so Muster erkennen, die einige Fragen schneller, präziser oder günstiger lösen können.	Die Datenbasis eines ADM-Prozessen kann Verzerrungen enthalten, die durch das Verfahren scheinbar objektiviert werden. Wenn die Kausalitäten hinter Korrelationen nicht überprüft werden, ist die Gefahr nicht beabsichtigter, aber billigend in Kauf genommener systematischer Diskriminierung groß.
Konsistenz der Anwendung	Algorithmenbasierte Prognosen arbeiten zuverlässig die vorgegebene Entscheidungslogik in jedem Einzelfall ab. Im Gegensatz zu menschlichen Entscheidern ist Software nicht tagesformabhängig und wendet nicht willkürlich in Einzelfällen neue, unter Umständen ungeeignete Kriterien an.	Bei ungewöhnlichen Fällen fehlt oft die Flexibilität, relevantes Unerwartetes auszuwerten und entsprechend zu reagieren. Auch Fehler in den Trainingsdaten oder der Entscheidungslogik wendet ein ADM-System auf jedem Einzelfall zuverlässig an.
Skalierbarkeit	Der Anwendungsbereich eingesetzter Software ist potenziell um ein Vielfaches größer als der Einflussbereich eines menschlichen Entscheiders, weil die Entscheidungslogik eines Systems sehr günstig auf nahezu unbegrenzt viele Fälle anwendbar ist.	Die leichte Skalierbarkeit von ADM-Prozessen kann dazu führen, dass die Vielfalt der eingesetzten und einsetzbaren ADM-Prozesse sinkt und dass viel häufiger und bei weit mehr Anlässen maschinell bewertet wird, als dies möglicherweise gesellschaftlich gewünscht ist.
Überprüfbarkeit	Datenbasierte und digitale Systeme können so gestaltet werden, dass Erklärbarkeit, Nachvollziehbarkeit, unabhängige Überprüfbarkeit und die Möglichkeiten zur forensischen Datenanalyse gegeben sind.	Die Begründung von Entscheidungen sowie eine unabhängige Evaluation sind in vielen Fällen qua Design und Betreibermodellen nur eingeschränkt institutionalisiert, möglich oder verständlich.
Anpassungsfähigkeit	ADM-Prozesse können neuen Umständen angepasst werden – entweder durch neue Trainingsdaten oder durch selbstlernend angelegte Systeme.	Die Symmetrie der Anpassungsfähigkeit in alle Richtungen ist abhängig von der Gestaltung des Prozesses. Auch einseitige Anpassung ist möglich.

Effizienz	Die maschinelle Auswertung großer Datenmengen ist in der Regel günstiger als die Auswertung vergleichbarer Datenmengen durch menschliche Analysten.	Effizienzgewinne durch ADM-Prozesse können verdecken, dass die absolut zur Verfügung stehenden Mittel insgesamt zu gering oder inadäquat sind.
Personalisierung	ADM-Prozesse können den Zugang zu personalisierten Angeboten und Dienstleistungen demokratisieren, die aufgrund der Kosten bislang wenigen vorbehalten waren. Zum Beispiel: Für die Recherchebreite und -tiefe einer Suchmaschinenanfrage waren vor dem Web mehrere wissenschaftliche Assistenten und Bibliothekare nötig.	Wo ADM-Prozesse das Massengeschäft dominieren, kommen nur wenige Privilegierte in den Genuss, von Menschen bewertet zu werden – was zum Beispiel bei der Bewerbervorauswahl oder Kreditvergabe in ungewöhnlichen Konstellationen Vorteile haben kann.
Menschliche Wahrnehmung maschineller Entscheidungen	ADM-Prozesse können konsistent zu statistischen Vorhersagen kommen. In bestimmten Fällen sind diese zuverlässiger als menschliche Experten – hier kann Software als Ergänzung mehr Zeit fürs Wesentliche schaffen.	Menschen können softwarebasierte Prognosen als verlässlicher, objektiver und aussagekräftiger als andere Informationen empfinden. Das kann dazu führen, dass Menschen die Empfehlungen und Prognosen der Software im Einzelfall nicht hinterfragen oder davon nur weiter in die vorgeschlagene Richtung abweichen.

4 Literatur

Algorithmwatch (2016). „Das ADM-Manifest“. https://algorithmwatch.org/das-adm-manifest-the-adm-manifesto/ (Download 19.2.2017).

Angwin, Julia, Lauren Kirchner, Jeff Larson und Surya Mattu (2016). „Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And it’s Biased Against Blacks“. 23.5.2015. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Download 11.12.2016).

Avery, Robert B., Kenneth P. Brevoort und Glenn Canner (2012). „Does Credit Scoring Produce a Disparate Impact?“. *Real Estate Economics* 40 s1. S65–S114.

Barry-Jester, Anna M., Ben Casselman und Donna Goldstein (2015). „The New Science of Sentencing“. *The Marshall Project.* 4.4.2015. https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing#.xXEp6R5rD (Download 24.4.2017).

Baumgartner, Fabian (2015). „Deutliche Zunahme im Jahr 2015: Wieder mehr Einbrüche in der Stadt Zürich“. *Neue Zürcher Zeitung* 27.10.2015. http://www.nzz.ch/zuerich/stadt-zuerich/wieder-mehr-einbrueche-in-der-stadt-zuerich-1.18636278 (Download 24.4.2017).

Berne, Xavier (2016). „Pour dévoiler l’algorithme d’Admission Post-Bac, l’Éducation nationale opte pour le papier“. 19.10.2015. https://www.nextinpact.com/news/101809-pour-devoiler-l-algorithme-d-admission-post-bac-l-education-nationale-opte-pour-papier.htm (Download 8.2.2017).

Brühl, Jannis (2014). „Ermitteln mit ‚Predictive Policing‘-Algorithmen: Polizei-Software soll Verbrechen voraussagen“. *Süddeutsche Zeitung* 12.9.2014. http://www.sueddeutsche.de/digital/ermitteln-mit-predictive-policing-algorithmen-polizei-software-soll-die-zukunft-voraussagen-1.2121942 (Download 24.4.2017).

Brühl, Jannis, und Florian Fuchs (2014). „Polizei-Software zur Vorhersage von Verbrechen: Gesucht: Einbrecher der Zukunft“. *Süddeutsche Zeitung* 12.9.2014. http://www.sueddeutsche.de/digital/polizei-software-zur-vorhersage-von-verbrechen-gesucht-einbrecher-der-zukunft-1.2115086 (Download 24.4.2017).

Bundesgerichtshof (2014). „Urteil des VI. Zivilsenats vom 28.1.2014 – VI ZR 156/13“. 28.1.2014. http://juris.bundesgerichtshof.de/cgi-bin/rechtsprechung/document.py?Gericht=bgh&Art=en&sid=6e60182ecc6717735edc41833c989561&nr=66910&pos=0&anz=1 (Download 24.4.2017).

Bundespolizei (2015). „EasyPASS“. http://www.easypass.de/EasyPass/DE/EasyPASS-RTP/rtp_node.html (Download 19.2.2017).

Chicago Department of Public Health (2016). „Fifty Years Fighting Lead in Chicago The Plan for a Lead Free Generation“. 5.7.2016. https://www.cityofchicago.org/content/dam/city/depts/cdph/food_env/general/Lead_Poison_Prevention_Program/CDPH_LeadBrochure_10172016.pdf (Download 24.4.2017).

Christin, Angele, Alex Rosenblat und Danah Boyd (2015). „Courts and Predictive Algorithms“. *Data & CivilRight: Criminal Justice and Civil Rights Primer*. 27.10.2015. http://www.datacivilrights.org/pubs/2015-1027/Courts_and_Predictive_Algorithms.pdf (Download 24.4.2017).

Citron, Danielle Keats, und Frank A. Pasquale (2014). „The Scored Society: Due Process for Automated Predictions“. *Washington Law Review* (89) 1. 1–33.

Consumer Reports (2015). „How a Credit Score Affects Your Car Insurance“. http://www.consumerreports.org/cro/car-insurance/credit-scores-affect-auto-insurance-rates/index.htm#creditmap (Download 24.4.2017).

Danziger, Shai, Jonathan Levav und Liora Avnaim-Pesso (2011). „Extraneous factors in judicial decisions“. *Proceedings of the National Academy of Sciences* (108) 17. 6889–6892. https://doi.org/10.1073/pnas.1018033108 (Download 24.4.2017).

Davey, Monica (2016). „Chicago Police Try to Predict Who May Shoot or Be Shot“. *The New York Times* 23.5.2016. http://www.nytimes.com/2016/05/24/us/armed-with-data-chicago-police-try-to-predict-who-may-shoot-or-be-shot.html (Download 24.4.2017).

de Coustin, Paul (2016). „APB: les explications du ministère ne lèvent pas tous les doutes“. *Le Figaro* 6.2.2016. http://etudiant.lefigaro.fr/les-news/actu/detail/article/l-algorithme-d-admission-post-bac-se-devoile-20621/ (Download 8.2.2017).

Desmarais, Sarah L., und Jay P. Singh (2013). „Risk Assessment Instruments Validated and Implemented in Correctional Settings in the United States“. Lexington, KY: Council of State Governments. https://csgjusticecenter.org/wp-content/uploads/2014/07/Risk-Assessment-Instruments-Validated-and-Implemented-in-Correctional-Settings-in-the-United-States.pdf (Download 24.4.2017).

Dunleavy, Eric M. (2016). „Written Testimony of Eric M. Dunleavy, PhD, Director“. Washington D.C. https://www.eeoc.gov/eeoc/meetings/10-13-16/dunleavy.cfm (Download 24.4.2017).

Eckhardt, Andres, Tim Weitzel, Sven Laumer, Christian Maier, Caroline Oehlhorn, Jakob Wirth und Christoph Weinert (2016). „Techniksprung in der Rekrutierung“. https://www.uni-bamberg.de/fileadmin/uni/fakultaeten/wiai_lehrstuehle/isdl/Recruiting_Trends_2016_-_Techniksprung_in_der_Rekrutierung_v_WEB.PDF (Download 24.4.2017).

Fair Isaac Corporation (2017). „How FICO Credit Score is Calculated“. http://www.myfico.com/crediteducation/whatsinyourscore.aspx (Download 30.1.2017).

Federal Trade Commission et al. (2012). „Report to Congress Under Section 319 of the Fair and Accurate Credit Transactions Act of 2003“. December.

Felten, Ed, National Science and Technology Council und Committee on Technology (2016). „Preparing for the Future of Artificial Intelligence“. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf (Download 24.4.2017).

Ferguson, Andrew Guthrie (2012). „Predictive Policing and Reasonable Suspicion“. *Emory Law Journal* 259. https://papers.ssrn.com/abstract=2050001 (Download 24.4.2017).

Flores, Anthony W., Kristin Bechtel und Christopher T. Lowenkamp (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'“. *Unpublished manuscript*. https://www.researchgate.net/profile/Christopher_Lowenkamp/publication/306032039_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/links/57ab619908ae42ba52aedbab.pdf (Download 24.4.2017).

Frouillou, Leila (2016). „Post-bac admission: an algorithmically constrained ‚free choice'“. http://www.jssj.org/wp-content/uploads/2016/07/JSSJ10_3_VA.pdf (Download 24.4.2017).

„Further Education in France“. *Angloinfo*. http://www.angloinfo.com/how-to/france/family/schooling-education/further-education (Download 8.2.2017).

Garvie, Clare, Alvaro M. Bedoya und Jonathan Frankle (2016). *The Perpetual Line-Up*. Washington D.C.: Center on Privacy and Technology at Georgetown Law. https://www.perpetuallineup.org/sites/default/files/2016-12/The%20Perpetual%20Line-Up%20-%20Center%20on%20Privacy%20and%20Technology%20at%20Georgetown%20Law%20-%20121616.pdf (Download 24.4.2017).

Gluba, Alexander (2014). *Predictive Policing – eine Bestandsaufnahme*. LKA Niedersachsen: Hannover. https://netzpolitik.org/wp-upload/LKA_NRW_Predictive_Policing.pdf (Download 24.4.2017).

Graveleau, Séverin (2016, Juni 1). „Admission post-bac, l'algorithme révélateur des failles de l'université". *Le Monde.fr* 1.6.2016. http://www.lemonde.fr/campus/article/2016/06/01/admission-post-bac-l-algorithme-revelateur-des-failles-de-l-universite_4929949_4401467.html (Download 24.4.2017).

Hannah-Moffat, Kelly Hannah, Paula Maurutto und Sarah Turnbull (2009). „Negotiated Risk: Actuarial Illusions and Discretion in Probation". *Canadian Journal of Law and Society* (24) 03. 391–409. https://doi.org/10.1017/S0829320100010097 (Download 24.4.2017).

Hawthorne, Michael (2015). „Could Chicago prevent childhood lead poisoning before it happens?". *Chicago Tribune* 16.7.2015. http://www.chicagotribune.com/news/ct-lead-poisoning-solutions-20150707-story.html (Download 3.1.2017).

Horton, Michelle (2016). „Stanford scientists combine satellite data, machine learning to map poverty". *Stanford News Service* 18.8.2016. http://news.stanford.edu/press-releases/2016/08/18/combining-satellg-to-map-poverty/ (Download 3.1.2017).

Hunt, Priscilla, Jessica M. Saunders und John S. Hollywood (2014). *Evaluation of the Shreveport predictive policing experiment*. Santa Monica CA: RAND Corporation.

Hurley, Mikella, und Julius Adebayo (2016). „Credit Scoring in the Era of Big Data". *Yale JL & Tech.* 18. 148–275.

Institut für musterbasierte Prognosetechnik (2014). „Near Repeat Prediction". http://www.ifmpt.de/prognostik/ (Download 9.1.2017).

Jean, Neil, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell und Stefano Ermon (2016). „Combining satellite imagery and machine learning to predict poverty". *Science*" (353) 6301. 790–794.

Johnson, Eddie T. (2016, Juli 14). „Special Order S09-11 Strategic Subject List (SSL) Dashboard". http://directives.chicagopolice.org/directives/data/a7a57b85-155e9f4b-50c15-5e9f-7742e3ac8b0ab2d3.html (Download 24.4.2017).

Jung, Dominik, Lorenz Kemper, Benedikt Kaempgen und Achim Rettinger (2015). „Predicting the Admission into Medical Studies in Germany: A Data Mining approach. Open Access at KIT". https://doi.org/10.5445/IR/1000045460 (Download 24.4.2017).

Koren, James Rufus (2015). „Some lenders are judging you on much more than finances". *Los Angeles Times* 19.12.2015. http://www.latimes.com/business/la-fi-new-credit-score-20151220-story.html (Download 24.4.2017).

Kurjuweit, Klaus (2017). „Berliner Bahnhof: Bahn testet intelligente Videoüberwachung am Südkreuz". *Der Tagesspiegel* 20.2.2017. http://www.tagesspiegel.de/berlin/berliner-bahnhof-bahn-testet-intelligente-videoueberwachung-am-suedkreuz/19413266.html (Download 21.2.2017).

Lefauconnier, Natacha (2016). „Leïla Frouillou : ‚APB promeut un libre choix d'études tout en étant socialement inégalitaire'". *Educpros* 16.6.2016. http://www.letudiant.fr/educpros/entretiens/leila-frouillou-apb-promeut-un-libre-choix-d-etudes-tout-en-etant-socialement-inegalitaire.html (Download 9.2.2017).

Lischka, Konrad (2015). „Wie die KI-Debatte falsch läuft und wo Software heute teilautonom entscheidet". 14.6.2015. http://www.konradlischka.info/2015/06/blog/wie-die-ki-debatte-falsch-laeuft-und-was-software-heute-schon-autonom-entscheidet/ (Download 24.4.2017).

Mastrobuoni, Giovanni (2015). „Crime is terribly revealing: Information technology and police productivity". Unpublished Paper. http://cep.lse.ac.uk/conference_papers/01_10_2015/mastrobuoni.pdf (Download 24.4.2017).

Mohler, George O., Martin B. Short, Sean Malinowski, Mark Johnson, George E. Tita, Andrea L. Bertozzi P. Jeff Brantingham (2015). „Randomized controlled field trials of predictive policing". *Journal of the American Statistical Association* (110) 512. 1399–1411. https://doi.org/http://dx.doi.org/10.1080/01621459.2015.1077710 (Download 24.4.2017).

Morgeson, Frederik P., Michael L. Campion, Robert L. Dipboye, John R. Hollenbeck, Kevin Murphy und Neal Schmitt (2007). „Reconsidering the use of personality tests in personnel selection contexts". *Personnel psychology* (60) 3. 683–729.

Morrison, Lennox (2017). „Speech analysis could now land you a promotion". http://www.bbc.com/capital/story/20170108-speech-analysis-could-now-land-you-a-promotion (Download 20.1.2017).

Northpointe (2015). „Practitioners Guide to COMPAS Core". http://images.google.de/imgres (Download 24.4.2017).

O'Neil, Cathy (2016a). *Weapons of math destruction: how big data increases inequality and threatens democracy* (First edition). New York: Crown.

O'Neil, Cathy (2016b). *Weapons of math destruction: how big data increases inequality and threatens democracy.* 1. Auflage. New York NY: Crown.

O'Neil, Cathy (2016c). „How algorithms rule our working lives". *The Guardian* 1.9.2016. https://www.theguardian.com/science/2016/sep/01/how-algorithms-rule-our-working-lives (Download 24.4.2017).

Patel, Prachi (2016). „Fighting Poverty With Satellite Images and Machine-Learning Wizardry". 18.8.2016. http://spectrum.ieee.org/tech-talk/aerospace/satellites/fighting-poverty-with-satellite-data-and-machine-learning-wizardry (Download 2.1.2017).

Pennsylvania Commission on Sentencing (2016). „Risk Assessment Project". http://pcs.la.psu.edu/publications-and-research/research-and-evaluation-reports/risk-assessment (Download 14.12.2016).

Perkowitz, Sidney (2016). „Should we trust predictive policing software to cut crime?". 27.10.2016. https://aeon.co/essays/should-we-trust-predictive-policing-software-to-cut-crime (Download 12.1.2017).

Pilpul, Martin (2016). „Wo Predictive Policing eingesetzt wird". Dezember 2016. https://blog.pilpul.me/wo-predictive-policing-eingesetzt-wird/ (Download 24.4.2017).

Plass-Fleßenkämper, Benedikt (2016). „Automatische Gesichtserkennung gegen den Terror – kann das funktionieren?". *wired.de* 25.8.2016. https://www.wired.de/collection/tech/automatische-gesichtserkennung-gegen-den-terror-kann-das-funktionieren (Download 12.2.2017).

Potash, Eric, Joe Brew, Alexander Loewi, Subhanrata Majumdar, Andrew Reece, Joe Walsh, Eric Rozier, Emile Jorgenson, Read Mansour und Rayid Ghani (2015). „Predictive Modeling for Public Health: Preventing Childhood Lead Poisoning". *KDD '15 Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 2039–2047. https://doi.org/10.1145/2783258.2788629 (Download 24.4.2017).

Prabhu, Robindra (2015). *Predictive policing – Can data analysis help the police to be in the right place at the right time?* Oslo: Teknologirådet. https://teknologiradet.no/wp-content/uploads/sites/19/2013/08/Predicitve-policing.pdf (Download 24.4.2017).

Revell, Timothy (2016). „Concerns as face recognition tech used to ‚identify' criminals". *New Scientist* 12.1.2016. https://www.newscientist.com/article/2114900-concerns-as-face-recognition-tech-used-to-identify-criminals/ (Download 24.4.2017).

Rivlin, Gary (2013). „Employers Pull Applicants' Credit Reports". *The New York Times* 11.5.2013. http://www.nytimes.com/2013/05/12/business/employers-pull-applicants-credit-reports.html (Download 24.4.2017).

Robinson, David, und Logan Koepke (2016). „Stuck in a Pattern – Early evidence on „predictive policing" and civil rights". *Upturn* August 2016. https://www.teamupturn.com/reports/2016/stuck-in-a-pattern (Download 24.4.2017).

Robinson, David, und Harlan Yu (2014). „Knowing the Score: New Data, Underwriting, and Marketing in the Consumer Credit Marketplace". https://www.teamupturn.com/static/files/Knowing_the_Score_Oct_2014_v1_1.pdf (Download 24.4.2017).

Saunders, Jessica (2016). „Pitfalls of Predictive Policing". *RAND*. http://www.rand.org/blog/2016/10/pitfalls-of-predictive-policing.html (Download 9.12.2016).

Saunders, Jessica, Priscilla Hunt und John S. Hollywood (2016). „Predictions put into practice: a quasi-experimental evaluation of Chicago's predictive policing pilot".*Journal of Experimental Criminology* (12) Sept. 347–371. https://doi.org/10.1007/s11292-016-9272-0 (Download 24.4.2017).

Schindler, Jessica, und Wolf Wiedmann-Schmidt (2015). „Kriminalität: Im roten Bereich". *Der Spiegel* 10/2015. http://www.spiegel.de/spiegel/print/d-132040367.html (Download 24.4.2017).

Schneider, Jan, Ruta Yemane und Martin Weinmann (2014). „Diskriminierung am Ausbildungsmarkt: Ausmaß, Ursachen und Handlungsperspektiven". Berlin: Forschungsbereich beim Sachverständigenrat deutscher Stiftungen für Integration und Migration (SVR). http://www.svr-migration.de/wp-content/uploads/2014/11/SVR-FB_Diskriminierung-am-Ausbildungsmarkt.pdf (Download 24.4.2017).

„Schufa-Klägerin zieht vor Verfassungsgericht". *Spiegel online* 4.11.2014. http://www.spiegel.de/wirtschaft/soziales/schufa-klaegerin-reicht-vor-verfassungsgericht-beschwerde-ein-a-964030.html (Download 24.4.2017).

Scoring (2010). „§ 28b Bundesdatenschutzgesetz Dritter Abschnitt – Datenverarbeitung nicht-öffentlicher Stellen und öffentlich-rechtlicher Wettbewerbsunternehmen (§§ 27–38a), Erster Unterabschnitt – Rechtsgrundlagen der Datenverarbeitung (§§ 27–32)".

Selbst, Andrew D. (2016). *Disparate Impact in Big Data Policing.* (SSRN Scholarly Paper No. ID 2819182). Rochester NY: Social Science Research Network. http://papers.ssrn.com/abstract=2819182 (Download 24.4.2017).

Smedley, Tim (2014). „Forget the CV, data decide careers". *Financial Times* 9.7.2014. https://www.ft.com/content/e3561cd0-dd11-11e3-8546-00144feabdc0 (Download 24.4.2017).

Society for Human Resource Management (2012). „SHRM Survey Findings: Background Checking – The Use of Credit Background Checks in Hiring Decisions". https://perma.cc/MMG9-QF4M (Download 24.4.2017).

Steinhart, David (2006). *Juvenile detention risk assessment: A practice guide to juvenile detention reform*. Band 1. Baltimore MD: Annie E Casey Foundation.

Stromboni, Camille (2017). „APB : le gouvernement recule sur le tirage au sort à l'entrée à l'université". *Le Monde.fr*. 18.1.2017. http://www.lemonde.fr/campus/article/2017/01/18/apb-le-gouvernement-recule-sur-le-tirage-au-sort-a-l-entree-a-l-universite_5064779_4401467.html (Download 24.4.2017).

„Terrorbekämpfung: De Maizière will Gesichtserkennung und Rucksackverbote". *Die Zeit* 21.8.2016. http://www.zeit.de/politik/deutschland/2016-08/terrorbekaempfung-thomas-de-maiziere-gesichtserkennung-flug-haefen (Download 24.4.2017).

The Demographic and Health Surveys Program (2014). „Wealth Index Construction". http://www.dhsprogram.com/topics/wealth-index/Wealth-Index-Construction.cfm (Download 3.1.2017).

The Leadership Conference on Civil and Human Rights, American Civil Liberties Union, Brennan Center for Justice, Center for Democracy, Technology, Center for Media Justice, Color of Change, Data&Society, Demand Progress, Electronic Frontier Foundation, freepress, media mobilizing project, 18MR.org, National Hispanic Media Coalition (NHMC), OpenMIC, Open Technology Institute und Public Knowledge (2016). „Predictive Policing Today: A Shared Statement of Civil Rights Concern". 31.8.2016. http://civilrightsdocs.info/pdf/FINAL_JointStatementPredictivePolicing.pdf (Download 24.4.2017).

Thompson, Madeleine (2016). „The French Educational Algorithm of Inefficiency". *Brown Political Reviev* 11.8.2016. http://www.brownpoliticalreview.org/2016/11/french-educational-algorithm/ (Download 24.4.2017).

Traub, Amy (2013). „Discredited: How employment credit checks keep qualified workers out of a job". *Demos* 7.

Trindel, Kelly (2016). „Written Testimony of Kelly Trindel". Washington DC. https://www.eeoc.gov/eeoc/meetings/10-13-16/trindel.cfm#fn6. (Download 24.4.2017).

Turner, Michael A., Patrick D. Walker, Chaudhuri Sukanya und Robin Varghese (2012). *A New Pathway to Financial Inclusion: Alternative Data, Credit Building, and Responsible Lending in the Wake of the Great Recession*. Durham NC: Politicy & Economicc Research Council.

United States Government Accountability Office (2016). „FACE RECOGNITION TECHNOLOGY FBI Should Better Ensure Privacy and Accuracy". (No. GAO-16-267). http://www.gao.gov/products/GAO-16-267 (Download 24.4.2017).

Vantagescore (2013). „What influences your VantageScore Credit Score?". https://www.vantagescore.com/pdf/VantageScore%20Infographic%2005.pdf (Download 24.4.2017).

Weber, Lauren, und Elizabeth Dwoskin (2014). „Are Workplace Personality Tests Fair?". *Wall Street Journal* 30.9.2014. http://www.wsj.com/articles/are-workplace-personality-tests-fair-1412044257 (Download 24.4.2017).

Weltbank (2015). „FAQs: Global Poverty Line Update". http://www.worldbank.org/en/topic/poverty/brief/global-poverty-line-faq (Download 3.1.2017).

Zweig, Katharina Anna (2016). „2. Arbeitspapier: Überprüfbarkeit von Algorithmen". 7.7.2016. http://algorithmwatch.org/zweites-arbeitspapier-ueberpruefbarkeit-algorithmen/ (Download 24.4.2017).

5 Executive Summary

Processes of algorithmic decision-making (ADM) now evaluate people in many areas of life. ADM processes have been used for years to categorize people, without any real discussion of whether those processes are fair or how they can be explained, verified or corrected. One potential reason for this is that the systems have little to do with artificial intelligence (AI) as it appears in science fiction. People often associate AI with qualities exhibited by fictional characters like HAL 9000 or Wintermute: intentionality and consciousness. Yet, until now, powerful AIs of this sort have only been found in literary works and films, and have nothing to do with the systems presented in this collection of case studies. The latter, however, already play a significant role in deciding legal matters, approving loans, admitting students to university, determining where and when police officers are on duty, calculating insurance rates and assisting customers who call service centers. All are programs which are specially designed to address specific problems and which impact the lives of many people. It's not about the future according to science fiction, it's about everyday reality today.

The nine case studies presented in this working paper demonstrate the opportunities and risks associated with such processes. The journey begins in the United States, with applications that we found only there, such as algorithms used in the criminal justice system to predict recidivism. Software-assisted pattern recognition, moreover, can help predict the risk of lead poisoning among children, depending on where they live. One transnational example illustrates some of the opportunities ADM offers: an artificial neuronal network that uses satellite photos to determine the regional distribution of poverty in developing countries almost as accurately as considerably more expensive on-site surveys. The results could be used to combat poverty by targeting those areas in the most distress and where, consequently, assistance can have the greatest impact. An example from France (university admissions) and several processes used in the US and subsequently adopted by Germany (e.g. location-specific predictive policing) show that the use of ADM processes is a global phenomenon, one that is also becoming more prevalent in Germany.

Every example highlights a typical problem and, with it, the need for a corrective response when designing future ADM processes meant to increase participation. To the greatest extent possible, the discussion here treats the problems as if they were discrete phenomena, knowing full well that the identified shortcomings often occur concurrently in practice.

To take advantage of the opportunities ADM offers in the area of participation, one overall goal must be set when ADM processes are planned, designed and implemented: ensuring that participation actually increases. If this is not the case, the use of these tools could in fact lead to greater social inequality. The risks and unwanted consequences seen in the chosen examples illustrate where corrective responses are required. Numerous potential problems can often be observed in the individual application scenarios. In each example given in this working paper, we highlight a typical response that should be considered when ADM processes designed to increase participation are developed.

Table 3: Corrective responses to ADM processes

Response	Description	Example
Ensure falsifiability	ADM processes can learn asymmetrically from mistakes. "Asymmetric" means that the system, by virtue of the design of the overall process, can only recognize in retrospect certain types of its own predictions which proved incorrect. When algorithms learn asymmetrically, the danger always exists that self-reinforcing feedback loops will occur.	Recidivism predictions used in the legal system

Ensure proper use	Institutional logic can lead to ADM processes being used for completely different purposes than originally envisioned by their developers. Such inappropriate uses must be avoided.	Predicting individual criminal behavior
Identify appropriate logic model for social impact	Algorithm-driven efficiency gains in individual process steps can obscure the question of whether the means used to solve a social problem are generally appropriate.	Predicting lead poisoning
Make concepts properly measurable	Social phenomena or issues such as poverty and social inequality are often hard to operationalize. Robust benchmarks developed through public discussion are therefore helpful.	Predicting patterns of poverty
Ensure comprehensive evaluation	The normative power of what is technically feasible all too easily eclipses the discussion of what makes sense from a social point of view. For example, the scalability of machine-based decisions can quickly lead to situations in which the appropriateness and consequences for society of using ADM processes have neither been debated nor verified.	Automatic face-recognition systems
Ensure diversity of ADM processes	Once developed, the decision-making logic behind an ADM process can be applied in a great number of instances without any substantial increase in cost. One result is that a limited number of ADM processes can predominate in certain areas of application. The more extensive the reach, the more difficult it is for individuals to escape the process or its consequences.	Preselection of candidates using online personality tests
Facilitate verifiability	Frequently, no effort is made to determine if an ADM process is sufficiently fair. Doing so is even impossible if the logic and nature of an algorithm is kept secret. Without verification by independent third parties, no informed debate on the opportunities and risks of a specific ADM process can take place.	University admissions in France
Consider social interdependencies	Even when use is very limited, the interdependences between ADM processes and their environment are highly complex. Only an analysis of the entire socio-informatic process can reveal the relationship between opportunities and risks.	Location-specific predictions of criminal behavior
Prevent misuse	Easily accessible predictions such as scoring results can be used for inappropriate purposes. Such misuse must be prevented at all costs.	Credit scoring in the US

This publication documents the preliminary results of our investigation of the topic. We are publishing it as a working paper to contribute to this rapidly developing field in a way that others can build on. We are therefore making it available as a working paper using a free license (CC BY-SA 3.0 DE), so that it might serve as the basis for discussion in workshops or during other considerations of the topic.

The case studies show how ADM processes influence decisions made about people. When machines evaluate us and when their predictions – as used in the legal system or by law enforcement officials – affect personal rights or – as is the case when candidates are being selected or credit assessed – issues of equality, then society must discuss the fairness of these processes and their impact on participation.

This is where a close look must be taken at the specific context, since not all ADM processes are equally risky. What society demands of ADM processes can vary depending on the consequences these processes have for

society as a whole or for individuals and their basic rights. Spelling-correction programs and navigation systems have a different impact on a person's life than processes which flag a person as being a credit risk or likely to commit a crime.

In sum, the opportunities and risks in the examples presented here point to a number of general factors related to ADM processes that can critically affect participation. These factors involve different aspects of the overall socio-informatic process and can be found on different levels. Here are three examples:

- **Shaping ADM processes on the micro and macro level:** Choosing data and setting criteria at the start of a development process can themselves reflect normative principles which sometimes touch on fundamental social issues.

- **Structure of suppliers and operators on the macro level**: Having a range of ADM processes and operators can increase participation (e.g. through credit assessments of people who have not been part of the system in the past), can make it easier to avoid the ADM process and can expand possibilities for falsification. Conversely, monopolistic structures increase the risk that individuals will "fall out of the system" and get left behind.

- **Use of ADM forecasts on the micro, meso and macro level**: The interplay of technology, society and individuals has a major impact on how and when algorithms are used and the influence they thus have. Key questions that must therefore be asked are: How do people (ADM developers and users, and the general public) deal with automated predictions? Do the processes include the possibility of challenging ADM results?

What are needed here are additional systematic analyses of the potential shortcomings of ADM processes on different levels – from the definition of the goals and the efforts to measure the issues at hand, to data collection, the selecting of algorithms and the embedding of processes in the relevant social context. Criteria are needed for determining the benefits of ADM processes on all levels and in all steps. The responses discussed here can provide initial impetus for addressing these issues (see Table 1: Corrective responses to ADM processes).

Table 4: Summary of opportunities and risks found in case studies

Dimension	Opportunities	Risks
Normative principles	When an ADM process is designed, normative decisions (e.g. about fairness criteria) must be made before the process is used. This offers an opportunity to discuss ethics issues thoroughly and publically at the very start and to document decisions.	ADM processes can contain hidden normative decisions. If discussion is only possible once the design phase is complete, any normative principles are more likely to be accepted as unalterable.
Data	Software can analyze a much greater volume of data than humans can, thereby identifying patterns and answering certain questions faster, more precisely and less expensively.	The data used for an ADM process can contain distortions that are seemingly objectified by the process itself. If the causalities behind the correlations are not verified, there is a significant danger that unintentional, systematic discrimination will become an accepted part of the process.

Consistency of application	Algorithm-based predictions apply the predetermined decision-making logic to each individual case. In contrast to human decision makers, software does not have good and bad days and does not in some cases arbitrarily use new, sometimes inappropriate criteria.	In exceptional cases, there is usually no possibility for assessing unexpected relevant events and reacting accordingly. ADM systems unfailingly make use of any incorrect training data and faulty decision-making logic.
Scalability	Software can be applied to an area of application that is potentially many times larger than what a human decision maker can respond to, since the decision-making logic used in a system can be applied at very low cost to a virtually limitless number of cases.	ADM processes are easily scalable, which can lead to a decrease in the range of such processes that are or can be used, and to machine-based decisions being made much more often and in many more instances that might be desirable from a societal point of view.
Verifiability	Data-driven and digital systems can be structured in a way that makes them clear and comprehensible, allows them to be explained and independently verified, and provides the possibility of forensic data analysis.	Because of process design and operational application, independent evaluations and explanations of decisions are often only possible, comprehensible or institutionalized to a limited degree.
Adaptability	ADM processes can be adapted to new conditions by using either new training data or self-learning systems.	The symmetry of the adaptability in all directions depends on how the process is designed. One-sided adaptation is also possible.
Efficiency	Having machines evaluate large amounts of data is usually cheaper than having human analysts evaluate the same amount.	Efficiency gains achieved through ADM processes can hide the fact that the absolute level of available resources is too low or inadequate.
Personalization	ADM processes can democratize access to personalized products and services that for cost-related reasons were previously only available to a limited number of people. For example, before the Internet, numerous research assistants and librarians were required to provide the breadth and depth of information that results from a single search-engine query.	When ADM processes are the main tools used for the mass market, only a privileged few have the opportunity to be evaluated by human decision makers, something that can be advantageous in non-standard situations when candidates are being preselected or credit scores awarded.
Human perception of machine-based decisions	ADM processes can be very consistent in making statistical predictions. In some cases, such predictions are more reliable than those made by human experts. This	People can view software-generated predictions as more reliable, objective and meaningful than other information. In some cases this can prevent people from questioning recommendations and predictions or

	means software can serve as a supplementary tool which frees up time for more important activities.	can result in their reacting to them only in the recommended manner.

An important – even definitive – quality factor must be stressed once again: An analysis of the opportunities, risks and societal consequences was only possible because independent third parties were able to verify the benefits of machine-based decisions. Institutions such as the investigative newsroom ProPublica, the US Government Accountability Office and the student-rights organization Droits des lycéens spent the time and financial resources needed to collect and evaluate data and to consider the relevant legal issues, allowing each of the algorithms to be explained and made transparent. Public debate on the impact of certain ADM processes thus depends completely on institutions of this sort – a situation that must change. It must be possible to verify and understand algorithmic decisions if an effective discussion is to take place, one which ensures that ADM processes actually increase participation and that machine-based decisions truly benefit people.

ADM processes will only contribute to the common good if they are discussed, criticized and corrected. We are still in a position to determine how we as a society want to make use of algorithms. We should not only consider how they are applied, but, in some cases, whether they should be used at all. For example, in those situations where society has chosen to promote solidarity and share risks, ADM processes cannot be permitted to individualize those risks. The guiding principle cannot be what is technically feasible, but what makes sense from a societal perspective – so that machine-based decisions truly do benefit people.

Adresse | Kontakt

Bertelsmann Stiftung
Carl-Bertelsmann-Straße 256
33311 Gütersloh
Telefon +49 5241 81-81216

Konrad Lischka
Taskforce Digitalisierung
Telefon +49 5241 81-81216
konrad.lischka@bertelsmann-stiftung.de

www.bertelsmann-stiftung.de

BertelsmannStiftung