Profiling, Prediction und Privatheit: Über das Verhältnis eines liberalen Privatheitbegriffs zu neueren Techniken der Verhaltensvorhersage

Martin Degeling

Big Data sorgt für Aufbruchsstimmung in Informatik und Wirtschaft. Informationen als Rohstoff des 21. Jahrhunderts versprechen Erkenntnis- und Marktanteilgewinne. Dabei werden z.B. Profile aus Informationen über Surf-, Shopping- und sonstige, messbare Verhaltensweisen erstellt und durch *Data Mining* ausgewertet. Automatisierte Analysen offerieren Zusammenfassungen, die Einschätzung über vergangenes und Voraussagen über zukünftiges Verhalten anbieten. Dabei bleibt es nicht bei einer reinen, rechnergestützten Beobachtung: *Behavioural targeting* zielt darauf, (Kauf-)Entscheidungen zu beeinflussen. Auf welchen technischen und normativen Grundlagen Anbieterinnen und Anbieter diese Dienstleistungen basieren und welche Strategien im Umgang mit ihnen bestehen, will der vorliegende Beitrag erhellen.

Big data is a source of huge optimism in the computing and the business sectors. Information is seen as an unlimited resource driving development in the twenty-first century. Profiles are generated and processed using information on surf histories, shopping and social networking activities. Automatic analyses offer summaries of past behaviour and anticipate future web-use. However algorithms do not just observe: behavioural targeting aims to influence (purchasing) decisions. The article analyses the legal and normative basis of such services and discusses strategies for dealing with them.

1 Einleitung

Unter dem Begriff *Big Data* wird seit einiger Zeit eine Entwicklung zusammengefasst, bei der die Informationen, die in unterschiedlichen Kontexten erhoben werden. Dazu gehören etwa Mess-, Nutzungs- oder Kundinnen- und Kundendaten (vgl. z.B. Mayer-Schönberger 2013). Diese Daten werden für einen anderen als ihren ursprünglichen Zweck nochmals verarbeitet und in ein Verhältnis gesetzt, um weitere Informationen zu generieren. Der Artikel diskutiert jenen Aspekt von *Big Data*, bei dem *Data Mining* und andere statistische Verfahren eingesetzt werden, um direkt oder indirekt Personen und deren Eigenschaften sowie zukünftige Verhaltensweisen von Gruppen von Menschen anhand berechenbarer Attribute zu beschreiben.

Dazu werden zuerst Beispiele aus dem Marketing- und Sicherheitsbereich beschrieben und ihr Einfluss auf die Privatheit von Einzelnen und Gruppen analysiert. Danach wird auf Konflikte der Methoden mit liberalen Privatheitsbegriffen eingegangen und in Bezug zur kybernetischen Hypothese diskutiert, die mit einer grundsätzlich verschiedenen Grundannahme die Basis der aktuellen Praxis erklärbar macht. Zuletzt werden rechtliche, normative und

informationstechnische Herangehensweisen vorgestellt, die sich mit den beschriebenen Verfahren auseinandersetzen.

Grundlegend für die Diskussion in diesem Artikel ist ein Verständnis von *Profiling* und *Data Mining* als sozio-technische Prozesse: Sowohl die technischen Aspekte, die Daten und Algorithmen, aber auch die damit verwobene soziale Praxis bei der Entwicklung und Nutzung der Technik sind zu betrachten.

Zur weiteren Einführung werden im Folgenden der Einsatz und die Folgen von *Scoring* und *Behavioural Targeting*, zwei konkrete Einsatzgebiete von *Data-Mining*, anhand von Beispielen veranschaulicht.

1.1 Risikobewertung mittels *Scoring*

Vielen Verbraucherinnen und Verbrauchern ist das Kredit-Scoring der Schutzgemeinschaft für allgemeine Kreditsicherung (SCHUFA) bekannt. Dort werden aus einer Menge an Einzelinformationen über Personen sogenannte *Scores* für die Kreditwürdigkeit errechnet – eine Art Risikoprofiling (vgl. Kamp/Weichert 2005). Scores sollen eine Aussage darüber treffen, wie wahrscheinlich es ist, dass eine Person eine Kreditrate, eine Wohnungsmiete oder eine Telefonrechnung nicht bezahlen wird. Die Einzelinformationen stammen dabei in der Regel von Banken und Vertriebspartnern der SCHUFA und umfassen neben Namen und Geburtsdatum etwa aktuelle und vorherige Wohnadressen, eine Liste aller Kreditverträge, Kontoeröffnungen, oft auch Telefonverträge; vor allem aber auch Informationen über offene Forderungen und Einträge über „abweichendes Zahlungsverhalten“ wie Zahlungsverzögerungen und -ausfälle. Der Score wird bei der SCHUFA vierteljährlich berechnet und anfragenden Unternehmen mitgeteilt, die dann damit unterschiedlich weiter verfahren. Je nach Geschäftskonzept entscheidet sich die Bank oder der Telefonanbieter, einen Vertrag einzugehen oder nicht. In der Regel fällt die endgültige Entscheidung nicht automatisiert, sondern durch eine Mitarbeiterin oder einen Mitarbeiter. Dabei ist es abhängig vom Unternehmen, das den Score bei der SCHUFA angefragt hat, ob die Entscheidung anhand der prozentualen Wahrscheinlichkeit, des absoluten Scores (ein Punktesystem) oder einer Klassifizierung auf einer Ordinalskala (sehr geringes Risiko bis sehr hohes Risiko) gefällt wird, die alle unterschiedlich berechnet werden.

Ähnliche Verfahren zur Klassifizierung von zukünftigem Verhalten werden

auch im Sicherheitsbereich eingesetzt. Insbesondere in der Flugsicherung werden personenbezogene Profile erstellt und das Sicherheitsrisiko aller Fluggäste vorab eingeschätzt, um die aufwendigen Einzelkontrollen an Flughäfen gezielter einzusetzen. Darunter fallen etwa das CAPPS II (*Computer Assisted Passenger Prescreening System II*) Verfahren, das in den USA u.a. auf Basis kommerzieller Datenbanken das Risiko der Reisenden bewertet und das nach starker Kritik eingestellt wurde (vgl. Barnett 2004). Das stattdessen aktuell eingesetzte *Secure Flight* Verfahren (vgl. Elias et al. 2005) stützt sich auf eine Reihe von Registern, die von (staatlichen) Sicherheitsbehörden geführt werden.

Auch bei der Berechnung von Kredit-Scores wird die Datenbasis, auf deren Grundlage die Scores berechnet werden, ständig erweitert. Neben den tatsächlichen Informationen über das Zahlungsverhalten werden soziodemografische Daten hinzugezogen wie Wohnort oder Alter (siehe Tabelle 1). Während ein Forschungsprojekt der SCHUFA, bei dem Möglichkeiten untersucht werden sollten, Daten aus *Social Networks* in die Score-Berechnung mit einzubeziehen, aufgrund öffentlicher Kritik abgesagt wurde, gehen kleinere Anbieter längst weiter. Das deutsche Unternehmen *Kreditech*[1], das in Polen, Russland und einigen anderen Ländern operiert, verleiht über das Internet kleinere Geldbeträge bis 200 Euro und verlangt dafür Einblick in die Profile von *Social Networks* wie *Facebook*, um über die Höhe des Kreditrahmens zu entscheiden. Dabei wird das gesamte Netz der Kontakte des Nutzers oder der Nutzerin analysiert und beispielsweise Informationen über Bildungsabschlüsse von Bekannten zusammen mit 8000[2] weiteren Datenpunkten zur Berechnung des Scores herangezogen.

Kategorie	**Beschreibung**
Datum	Datum der Berechnung des Scores (vierteljährlich)
Bezeichnung	z.B. Score für Banken, Telekommunikationsunternehmen, Versandhandel, Freiberufler

[1] Online verfügbar unter http://www.kreditech.com (zuletzt abgerufen am 25.06.2013).

[2] Eigenauskunft des Unternehmens.

Score-Wert	Numerischer Wert ~ 0 bis 10.000
Ratingstufe	A bis F
Erfüllungswahrscheinlichkeit	Prozentwert
Bisherige Zahlungsstörungen	– – bis ++ (deutlich überdurchschnittliches Risiko bis deutlich unterdurchschnittliches Risiko)
Kreditaktivität letztes Jahr	– – bis + +
Kreditnutzung	– – bis + +
Länge Kredithistorie	– – bis + +
Allgemeine Daten	– – bis + +
Anschriftendaten	– – bis + +
Bedeutung insgesamt	Sehr geringes bis sehr hohes Risiko

Tabelle 1: Elemente einer SCHUFA-Scorecard

Ziel des Scorings ist es, eine vergleichbare Messgröße für die (finanzielle) Vertrauenswürdigkeit zu schaffen, d.h. also Vertrauen dort herzustellen, wo hohe Risiken im Spiel sind oder es nicht effizient wäre sich ein persönliches Bild vom Geschäftspartner oder der Geschäftspartnerin zu machen. Der Score wird auf der Basis von Informationen berechnet, die als Spuren des Verhaltens einer Person dokumentiert sind, sowie aus Informationen über ihr Wohnumfeld. Das Ergebnis wird interpretiert als eine Aussage über die Wahrscheinlichkeit von zukünftigem Verhalten. Möglichkeiten, diese personenbezogenen Daten zu beeinflussen gibt es nur indirekt, indem man sich etwa möglichst so verhält, dass die aufgezeichneten Daten ein Bild erzeugen, das einen Score entstehen lässt, der der eigenen Selbstwahrnehmung der Vertrauenswürdigkeit bei Kreditrückzahlungen entspricht. Dabei sind die Bewertungen der Spuren allerdings meist wenig transparent und, wie etwa Wohnumfelddaten[3], nur bedingt kontrollier- und beeinflussbar. Die SCHUFA hat vor einigen Jahren in Gerichtsverfahren versucht, die Selbstauskunft von Bürgerinnen und Bürgern zu unterbinden, wollte also die berechnete Vertrauenswürdigkeit einer Person gegenüber dieser möglichst geheim halten. Am Ende konnte nur die Veröffentlichung der Berechnungslogik, die als Geschäftsgeheimnis gilt, durch

[3] Sie basieren selbst wiederum auf mittels statistischer Rechenverfahren erzeugter Vorstellungen davon, welches Wohnumfeld, welches Milieu, welche Altersgruppe, welche ethnische Herkunft vertrauenswürdiger oder weniger vertrauenswürdig erscheint.

die SCHUFA verhindert werden, eine Auskunft über den eigenen Score ist dagegen jährlich kostenlos möglich.

1.2 Beeinflussung durch *Behavioural Targeting*

Der Werbesektor ist ein weiterer Bereich, in dem *Data Mining* sehr häufig eingesetzt wird. Hier wird mittels großer Datenmengen versucht, Gruppen zu spezifizieren, um diesen *zielgenau* Produkte anbieten zu können – im Englischen *Targeting*.

Marketingabteilungen großer Handelsketten und Werbevermittler, online wie offline, bauen Datenbanken auf, in denen das Kaufverhalten der Kundinnen und Kunden gespeichert und analysiert wird. Ihre Methoden gehen dabei über die klassische Zielgruppenwerbung hinaus. In einem Artikel der *New York Times* aus dem vergangenen Jahr wird über die amerikanische Drogeriekette *Target* berichtet, die diese Daten für *predictive analytics* heranzieht (vgl. Duhigg 2012). Durch Informationen, die das Unternehmen mittels Kundenkarten über Einkäufe und Kreditkartenzahlungen erhebt, kann *Target* auf eine große Datenbasis zurückgreifen, die die hauseigene Statistikabteilung nutzen sollte, um die Zielgruppe ‚schwangere Frauen' zu identifizieren. Anhand einer Liste von 25 Produkten wurde ein *pregnancy prediction score* errechnet, der solche Kundinnen zu identifizieren versucht, die vermutlich gerade schwanger sind und daher vermeintlich bald eine große Menge an Babyprodukten benötigen.

Ähnlich operieren Werbetreibende im Internet. Beim sogenannten *Online (Behavioural* oder *Re-) Targeting* wird versucht, nicht nur anhand eines Profils personalisierte Werbung zu vermitteln, sondern auch die Reaktionen auf die Werbung zu erfassen und Informationen darüber wiederum für Anpassungen der Werbungsschaltungen zu nutzen, um so die Nutzerinnen und Nutzer zu einem bestimmten Verhalten zu (ver-)leiten. Dabei spielt die c*onversion rate* – d.h. die Anzahl der Verkäufe, Registrierungen oder einfach Klicks – pro Bannerwerbung eine zentrale Rolle. Mittels *Tracking*[4], meist durch Cookies (vgl. Mayer/Mitchell 2012), wird versucht, das Nutzungsverhalten im Internet möglichst lückenlos zu verfolgen. *Acxiom*, einer der größten Anbieter, beschreibt in einer Produktpräsentation folgendes Szenario[5]:

[4] Zur Aktualität des Tracking-Problems siehe u.a. International Working Group on Data Protection in Telecommunications (2013) oder Article 29 Data Protection Working Party (2010).

[5] Nach Singer (2012).

> Herr Higgs sieht bei *Facebook*, dass eine Freundin den Onlineshop *Bryce* ‚geliked' hat, was ihn dazu verleitet, sich die Webseite des Ladens anzuschauen und nach Druckern zu suchen, weil er vorhat, sich bald einen neuen zu besorgen. Da der Onlineshop mit Facebook verknüpft ist, wird sein Profil dort mit seinem Verhalten in Verbindung gebracht. Er registriert sich im Weiteren auf der Seite des Shops, kauft aber am Ende doch keinen Drucker. Als er am nächsten Tag auf einer Sportnachrichtenseite surft, wird sein Profil wiedererkannt und ihm wird Werbung für eben jene Drucker angezeigt, die er am Vortag nicht gekauft hat. Als er daraufhin wieder die Seite des Shops besucht, wird ihm vom System ein Rabatt angeboten, wenn er sich jetzt entscheiden würde. Der Rabatt erscheint nicht zufällig: Durch die Verknüpfung mit Facebook und das Wissen um die Sportseite, die Higgs angesurft hat, konnte *Acxiom* ihn in die Kategorie ‚cleverer Single' einsortieren – einer von 70 Clustern, denen unterschiedliche Profile zugeordnet werden. Das System weist ihm die Eigenschaften mobil, obere Mittelklasse, online Banker, sportbegeistert und preissensibel zu. Für dieses Cluster empfiehlt Acxiom einen Rabatt zur Förderung von Kaufentscheidungen.

Im Szenario ist Higgs begeistert von dem Angebot und kauft den Drucker. Das Szenario ist fiktiv und dient vor allem der Werbung für die Marketingfirmen. Es veranschaulicht gleichzeitig aber auch die Idee, die dahinter steht. Die Firmen versuchen gezielt, Werbung auf bestimmte Personen und Personengruppen zu lenken. Diesen folgen sie dann im Internet, um eine Anzeige, die ihrem Verhalten entspricht (*behavioural targeting*), möglichst häufig anzuzeigen (*re-targeting*) und etwa durch spezielle Preisgestaltung einen Kunden oder eine Kundin zu gewinnen.

Im Unterschied zu den vorher beschriebenen Scoring- und Profiling-Mechanismen ist das Online-Targeting wesentlich flexibler. Ziel ist es, möglichst viele Informationen über Nutzerinnen und Nutzer in Erfahrung zu bringen, um die Kategorisierung zu verfeinern. Jeder Kauf bzw. auch Nicht-Kauf fließt dabei zurück in die Algorithmen, die durch das Feedback optimiert werden. Das Ziel ist dabei die Beeinflussung der Nutzerinnen und Nutzer in ihren Kaufentscheidungen oder anderem vermarktbaren Verhalten. Der Unterschied zum Risiko-Scoring liegt aber in der tendenziellen Ignoranz gegenüber der tatsächlichen Identität des oder der Einzelnen, da die Information, in welche Kategorie jemand am besten passt, wichtiger ist als etwa der Name oder die Adresse. Stephan Noller, Gründer von *nugg.ad*, einer in Deutschland ansässigen Online-Werbe-Firma, ließ sich etwa mit dem Satz zitieren „Es geht nicht darum, den Leser zu finden, sondern nur sein statistisches Modell" (zit. nach Biermann 2010). Die Profile werden dabei zwar, wenn möglich, mit realen Personen verknüpft. Im Vordergrund stehen aber die Wiedererkennung eines vordefinierten Profils und die Beeinflussung des Verhaltens auf die Art, wie es für das jeweilige Profil vorgesehen ist. Zudem dient die Auswertung nicht als

Grundlage für die Entscheidung Dritter, wie etwa einer Bank, die über die Kreditvergabe entscheidet, sondern sie hat automatisierte Folgen über die angezeigte Werbung.

2 Exkurs zu den technischen Grundlagen von *Data Mining*

Bevor versucht wird, der Frage des Einflusses von Profiling auf Privatheit nachzugehen, folgt ein Exkurs zum Aufbau der technischen Systeme, die den beiden skizzierten Beispielen zugrunde liegen.

Die Berechnungen erfolgen in der Regel auf Basis von Methoden des *Data Mining*, d.h. der Analyse von – meist strukturierten – Datensätzen wie Tabellen mittels statistischer Verfahren, die mehr oder weniger lernend sind, d.h. sich rekursiv den beobachteten Daten anpassen (vgl. Fayyad et al. 1996). Wie sich bereits in den Beispielen zeigt, werden *Data-Mining*-Prozesse oft als Black Box dargestellt. Ein Algorithmus erhält einen Datensatz als Input, wie etwa Kredithistorie, und eine Zahl, die komplexe Informationen aggregiert (Score), erscheint als Output. Je nach Art des Inputs und des erwarteten Outputs sind aber einige Bedingungen vorab zu beachten, um z.B. einen Klassifizierer zu entwickeln, der Datensätze in Klassen wie hohes, normales oder geringes Risiko einteilt. Bei jedem Schritt müssen dabei Kompromisse eingegangen werden, die die Genauigkeit einschränken und deren Betrachtung für eine weitere Diskussion notwendig ist (vgl. Domingos 2012).

Zu Beginn müssen Daten erhoben werden: Merkmale von Personen, ihr Umfeld und ihr Verhalten müssen zu Informationen abstrahiert und digitalisiert werden. Oft wird dazu auf öffentliche Datenbanken wie die von Statistikämtern oder öffentlichen Befragungen zurückgegriffen (vgl. Tavani 1999). Die Erhebung erfolgt also nicht bei den Betroffenen selbst, sondern die Korrektheit der eingekauften Daten wird vorausgesetzt. Danach muss eine geeignete Form der *Datenrepräsentation* gefunden werden. Nicht alle Daten liegen klar strukturiert in Tabellen vor bzw. sind in solchen darstellbar. Bereits das Aufstellen von Kategorien in einer Tabelle bedeutet eine Filterung: Merkmale, die keinen Eingang in die Tabelle gefunden, haben können auch nicht mit anderen zusammen analysiert bzw. in Beziehung gesetzt werden. Auch werden durch das Erstellen einer Datenrepräsentation (als freier Text oder Zahl) bereits erste Kategorisierungen vorgenommen. Nur wenige Datenbanken erlauben z.B. eine

Differenzierung zwischen Geschlecht und Gender, geschweige denn eine andere Klassifizierung als die binäre in solchen Datenfeldern.

echnisch komplexer wird es bei der Analyse unstrukturierter Daten wie Fließtexten oder Bildern, bei denen der erste Schritt darin besteht, analysierbare Einheiten zu konstruieren – etwa Augen in einem Gesicht oder die grammatikalische Konstruktion eines Satzes zu erkennen. Zudem skalieren die existierenden Algorithmen, die die Daten auswerten, nicht beliebig. Das Hinzufügen weiterer Datenfelder (also etwa Anzahl der Attribute pro Datensatz oder Spalten pro Tabellenreihe) bedeutet häufig eine nicht-lineare Komplexitätssteigerung. Je mehr *Dimensionen* miteinander in Verbindung gesetzt werden, desto länger braucht der Computer zur Berechnung der Zusammenhänge.[6] Und auch wenn die verfügbare Rechenleistung stetig steigt, so erhöht eine größere Menge an Daten auch das Risiko, dass sich keine signifikanten Zusammenhänge mehr errechnen lassen. So analysiert die bereits erwähnte Drogeriemarktkette eine Liste von lediglich 25 Artikeln zu Ermittlung des *pregnancy prediction score*, anstatt die gesamte Angebotspalette in die Berechnung mit einzubeziehen. *Data Mining* setzt daher *domänenspezifisches Wissen* voraus: Wer die Produktpalette (*Domäne*) kennt, kann viele Artikel schon vor der ersten Berechnung ausschließen und so die mathematischen Verfahren vereinfachen. Gleichzeitig finden durch diese Vorauswahl Vorurteile Eingang in die Datenbanken und Analysen (*Bias*).

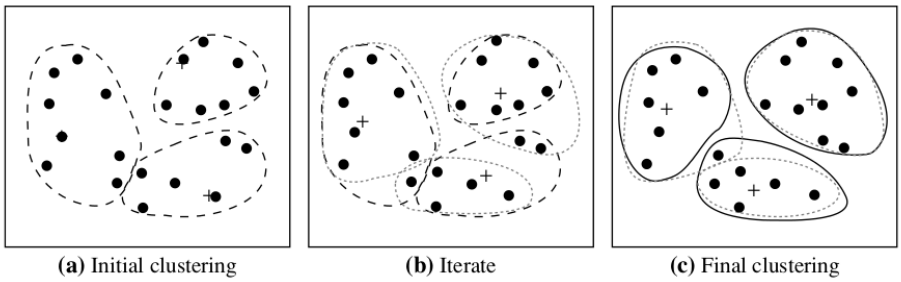

Abbildung 1: Darstellung eines mehrstufigen Clusterings von Daten mit zwei Dimensionen (Quelle: Han et al. 2011: 453).

Zur Entwicklung solcher Algorithmen wird in der Regel mit *Trainings-* und *Test-Datensätzen* gearbeitet. Trainingsdaten enthalten Informationen zu allen

[6] Zwar erlaubt das von Google entwickelte *Map-Reduce*-Verfahren eine wesentlich effizientere Verarbeitung von extrem großen Datenmengen (*Big Data*), erreicht dies aber vor allem durch die Parallelisierung von Operationen.

Dimensionen und auch zum gewünschten Ergebniswert (schwanger oder nicht). Dem Test-Datensatz wird diese Information für einen Test entfernt. Wurden mittels der Trainingsdaten Gewichtungen für die einzelnen Dimensionen errechnet, wird mittels einer Evaluierungsfunktion getestet, ob die Werte für diesen anderen Datensatz ebenfalls plausibel sind. So soll ein *Overfitting* verhindert werden, bei dem der Algorithmus für einen konkreten Trainingsdatensatz eine sehr hohe Genauigkeit hat, aber zu weiteren Datensätzen keine brauchbaren Ergebnisse liefert, weil die Zusammenhänge zu genau gelernt wurden. Im Gegensatz dazu entsteht eine zu große Generalisierung, wenn Datensätze automatisiert gruppiert (*Clustering*; siehe Abbildung 1) werden, etwa in ähnliche Verhaltensmuster: Ist ein Algorithmus zu generell, berechnet er nur eine Kategorie in die alle Datensätze einsortiert werden, sodass keine Unterscheidung möglich ist. Die Bewertung der Ergebnisse insbesondere teil-automatischer Methoden erfordert also menschliches Abwägen, etwa über die beste Anzahl von Kategorien.

Aber auch nachdem eine Datenmenge zufriedenstellend geordnet wurde, sind die Verfahren nicht abgeschlossen. Bei der SCHUFA werden z.B. Scores nach drei Monaten anhand der neu zur Verfügung stehenden Daten mit einem dann jeweils angepassten und neu gelernten Algorithmus neu berechnet. Beim *Online-Targeting* wiederum werden die Informationen bei jedem Schritt neu verarbeitet, d.h. der *Targeting-Algorithmus* tritt indirekt in Interaktion mit den Nutzerinnen und Nutzern, indem er ihr Feedback – also wie sie sich im Netz bewegen – aufnimmt und neu kalkuliert.

3 Analyse von *Data Mining* in der Praxis

3.1 Die Ökonomische Dimension: *Panoptic Sort*

Trotz angenommener Neutralität der auf Daten beruhenden, technischen Analyse geraten die Ergebnisse einer Berechnung gelegentlich in die Kritik, weil sich in der Praxis spezifische, kulturell und geschichtlich bedingte Vorurteile in *Data-Mining*-Verfahren reproduzieren. Zu Beginn des Jahrtausends wurde zum Beispiel bekannt, dass der mittlerweile insolvente Onlineversender *Kozmo* nur Bestellungen annahm, deren Lieferadresse sich nicht in solchen Stadtteilen befand, für die eine unterdurchschnittliche Rückzahlungsquote berechnet wurde (vgl. Danna/Gandy 2002). Eine Methode, die bereits in den 60er-Jahren in

den USA als *Redlining* in Verruf geraten war.[7]

Oscar Gandy (1993) hat für dieses und ähnliche Phänomene Anfang der 90er Jahre den Begriff des *panoptic sort* geprägt, in Anlehnung an jene Architekturform, die Michel Foucault in Anlehnung an Jeremy Bentham beschreibt, und die die Überwachung von Gefangenen als Dispositiv organisiert, das die Überwachung durch Menschen obsolet macht. Er beschreibt den

> panoptic sort as a kind of high-tech cybernetic triage through which individuals and groups of people are being sorted according to their presumed economic or political value. (Gandy 1993: 1)

Ähnlich dem Panoptikum, für das konstitutiv ist, dass Gefangene nicht wissen (müssen), ob sie beobachtet werden, verweist Gandys auf die ständige und unsichtbare Anwesenheit von Klassifizierungsmechanismen, die darüber entscheiden, welcher ökonomische Spielraum zugebilligt wird.[8] Ausbrechen könne aus dem *panoptic sort* nur, wer die persönliche Vergangenheit von der Verarbeitung ausschließen könne. Notwendig sei dazu das vollständige Löschen aller digitalen Spuren. Dabei stützt sich der *panoptic sort* auf die Datensammlungen sowohl von staatlicher Seite (in den USA öffentliche geführte Register etwa in Wählerinnen- und Wählerverzeichnissen oder Zensusdaten) als auch von privatwirtschaftlicher Seite (wie Marketingbefragungen und Adresslisten). Neben dem *Redlining*, das Personen vollständig vom Markt auszuschließen vermag, sieht Gandy für die verbleibenden Marktteilnehmerinnen und -teilnehmer das Problem einer vollständigen Preisdifferenzierung (vgl. Danna/Gandy 2002: 380). Darauf verweist das vorgestellte Szenario der Firma *Acxiom*, die potentielle Käuferinnen und Käufer mit Rabatten lockt, während diejenigen, die sich bereits für ein spezifisches Produkt entschieden haben (oder aber gar keine Alternative zu diesem haben) den höheren Normalpreis zahlen.

Dem vorurteilsgeprägten *Redlining* ähnelt das aktuelle Beispiel einer amerikanischen Personensuchmaschine[9], die Werbung für ihren Service bei

[7] Im Fall von *Kozmo* ist außerdem bemerkenswert, dass sich häufig die Lagerhallen, von denen aus die Auslieferung organisiert wurde, in eben jenen Stadtteilen befanden, da dort auch die Mieten günstig waren.

[8] Der Kurzschluss, der die Übertragung der Figur der Gefangenen auf die Figur der Marktteilnehmenden erlaubt, kann hier nicht abschließend diskutiert werden.

[9] Dabei handelt es sich um Webseiten, die öffentlich zugängliche Daten über Personen zusammenführen und als Dossiers anbieten. Während sich in Deutschland die Portale meist auf die Kombination von Telefonbucheinträgen und Webseiten beschränken, liegen den amerikanischen Varianten zusätzlich öffentliche Register wie Listen von verurteilten Straftäterinnen und Straftätern

Google schaltet. In einer Studie (Sweeney 2013) konnte nachgewiesen werden, dass die Werbetexte anhand der vermuteten Hautfarbe jener Person, nach der gesucht werden soll, adaptiert wurden. Suchte man etwa auf einer Webseite, die ihre Werbung über *Google* bezog, nach „Latanya Farell“ – ein Vorname der eher *black identifying* ist –, enthielten die Werbetexte signifikant häufiger tendenziöse Anspielungen wie „Latanya Farell, Arrested?“ Bei Namen die *white identifying* sind wurden stattdessen nur Phrasen wie „We found Kirsten Haring: 1) Contact Kirsten Haring 2) Current Phone, Adresse & More“ ausgegeben (Sweeney 2013). In beiden Fällen lagen dem beworbenen Service keine Einträge aus polizeilichen Datenbanken vor.

3.2 Datenschutz und anonymisierte Daten

Insbesondere dem Scoring wie es die SCHUFA betreibt sind in der Regel nicht nur rechtliche Grenzen gesetzt, sondern es ist sogar mit Pflichten verbunden. Neben dem seit 2001 geltenden Verbot der automatisierten Einzelentscheidung (§ 6a BDSG) ist für jede SCHUFA-Abfrage zwingend eine schriftliche Einwilligung vorzulegen. Zudem wurde die SCHUFA verpflichtet, unentgeltlich Auskunft zu erteilen (vgl. Weichert 2006).[10] Die SCHUFA eignete sich auch deswegen als Modellfall, da sie auf die Nutzung direkt personenbezogener Daten angewiesen ist wie sie vom Gesetz definiert werden. Zudem sind im Zusammenhang mit der SCHUFA viele Fälle falsch dokumentierter Daten bekannt geworden (vgl. Bäumler 2002).

Auch die Verfahren, die von Sicherheitsbehörden eingesetzt werden, sind zumindest Thema öffentlicher Debatten, die, wie im Fall von CAPPS II, auch dazu führen können, dass Programme eingestellt werden.[11] Die Kritik an den Verfahren ist auch aus Sicht des Datenschutzes lange bekannt (vgl. Konferenz der Datenschutzbeauftragten des Bundes und der Länder 2000; Tavani 1999). Wesentliche Argumente sind, dass durch das Zusammenführen unterschiedlicher Datenbanken gegen das Prinzip der Zweckbindung verstoßen

zugrunde.

[10] Bis 2001 floss die Tatsache, dass man eine Eigenauskunft eingeholt hat, noch negativ in den Gesamtscore ein. Eine Auskunft kostete bis vor wenigen Jahren 8 Euro. Mittlerweile gibt sich die SCHUFA offener, eine Eigenauskunft pro Jahr ist kostenlos.

[11] Wobei das Folgeprogramm *Secure Flight* sich vor allem auf Datenbanken mit geheimdienstlichen Quellen stützt, die sich jeglicher Kritik entziehen, selbst wenn, wie im Fall von Mikey Hicks (vgl. Alvarez 2010), Fehler nachgewiesen werden können.

wird. D.h., dass in eine solche Nutzung nicht eingewilligt wurde oder dass bei der *Knowledge discovery* Informationen ermittelt werden, die so nicht direkt bei den Betroffenen erhoben wurden. Viele Dienste, insbesondere jene, die zum Werbe-Targeting eingesetzt werden, fallen aber nicht unter den Anwendungsbereich eines Datenschutzgesetzes.[12] Sie arbeiten meist mit anonymen oder stark pseudonymisierten Daten, um Klassifizierungen zu entwickeln und sind damit, zumindest in Bezug auf das geltende Datenschutzrecht, unbedenklich. Auch die Nutzungsanalyse bei *Facebook* arbeitet etwa nach Unternehmensangaben auf Basis von „aggregated and anonymized clusters of users“ (Rodgers 2013) und somit nicht mit Datensätzen, die sich auf konkrete Personen beziehen. Im zuvor beschriebenen Beispiel ist es unwichtig, ob Herr Higgs oder Max Powers einen Drucker kaufen möchte. Die Werbung richtet sich an eine Person der Zielgruppe ‚Obere Mittelklasse‘ mit ‚Interesse an Sport‘.

Viel wichtiger als die Frage *wer* etwas kauft, ist für die Werbetreibenden die Information darüber *was* jemand kaufen will und kaufen könnte. *Amazons* Empfehlungsmechanismus etwa arbeitet vor allem mit so genannten *Item-to-Item* Relationen (vgl. Linden et al. 2003): Welche Produkte werden zusammen gekauft? Welche nacheinander? Konkret personenbezogene Informationen spielen erst beim Abschluss des Kaufs eine Rolle. Statt Kundinnen und Kunden bilden die Produkte die Referenzpunkte der Algorithmen. Auch wenn es also nicht um die konkrete einzelne Person geht, wirken ermittelte, vermeintliche Gruppenzugehörigkeiten auf uns zurück und beeinflussen unterschiedliche Dimensionen von Privatheit.

4 *Data Mining* und Privatheit

4.1 Dimensionen von Privatheit

Die beschriebenen Phänomene stehen aus vielerlei Gründen in einem direkten Zusammenhang mit Aspekten von Privatheit. Die folgende Analyse stützt sich dabei vor allem auf Rössler (2001), die den *Wert des Privaten* für eine liberale Gesellschaft als Ganzes vor allem aus dem Wert für die Autonomie der Einzelnen ableitet. Rössler unterscheidet dabei drei Dimensionen des Privaten, die lokale, die dezisionale und die informationelle Privatheit.

[12] Ein Faktor ist auch, dass viele der Anbieter nicht von Deutschland oder Europa aus operieren und damit nicht unter hiesige Gesetze fallen.

Der Bezug zwischen Profiling und Privatheit lässt sich zu zwei dieser Dimensionen herstellen. Erstens konfligieren die beschriebenen Systeme mit der informationellen Privatheit, die in der deutschen Rechtspraxis unter dem Begriff der informationellen Selbstbestimmung verhandelt wird. Die kategorisierenden Profiling-Mechanismen versuchen, möglichst viele Daten zu sammeln, zu aggregieren und weiterzugeben. Einen Überblick darüber zu bekommen, was andere potentiell oder tatsächlich über eine Person wissen und damit informationelle Selbstbestimmung auszuüben, ist unmöglich. Dabei ist

> der Schutz informationeller Privatheit [...] deshalb so wichtig für Personen, weil es für ihr Selbstverständnis als autonome Personen konstitutiv ist, (in ihr bekannten Grenzen) Kontrolle über ihre Selbstdarstellung zu haben, also Kontrolle darüber, wie sie sich wem gegenüber in welchen Kontexten präsentieren, inszenieren, geben wollen, als welche sie sich in welchen Kontexten verstehen und wie sie verstanden werden wollen. (Rössler 2001: 209)

Ein Scheitern ist dabei unumgänglich, wenn nicht klar ist, wie die Mechanismen der Datenverarbeitung funktionieren, welche Kategorien existieren in die man potentiell einsortiert werden könnte, und auf welcher Basis von Informationen (die ggf. bereitwillig preisgegeben wurden) weitere Informationen generiert werden. Gerade diese Abstraktion zweiter Ordnung ist charakteristisch für *Data Mining*. So werden den Userinnen und Usern Verhaltensweisen zugeschrieben, die sie (bisher) möglicherweise nicht gezeigt haben bzw. die aus einem den *Data Mining*-Praktiken eigentümlichen Begriff von Verhalten entspringen, deren bloße Annahme aber bereits Folgen nach sich ziehen kann. Eine klassische Verletzung informationeller Privatheit dokumentiert etwa das Beispiel der Schwangeren, deren Schwangerschaft die Firma *Target* errechnet hatte und – da in diesem Fall der Algorithmus sozusagen Recht behielt und das errechnete Verhalten mit dem tatsächlichen Verhalten der Kundin alsbald korrelieren sollte – mittels der eilig versendeten Werbung auch den Rest der Familie frühzeitig über die Schwangerschaft informierte und für Konflikte sorgte.

Darüber hinaus hat Profiling einen negativen Einfluss auf die dezisionale Privatheit, die Rössler wie folgt beschreibt:

> [W]enn man Selbstbestimmung, die Autonomie einer Person, auch so verstehen muss, dass eine Person das Recht hat, [...] die Autorin ihrer eigenen Biographie zu sein, dann muss dies in sozialen Kontexten auch bedeuten, dass das je eigene Leben nicht von solchen anderen Personen kommentiert oder interpretiert, oder auch stärker: beeinflusst wird, denen sie gerade keine solche Interpretationshoheit über ihre Leben zubilligen will. (Rössler 2001: 152)

Eben diese Beeinflussung findet insbesondere beim Online-Targeting statt. Anhand des Surfverhaltens werden Annahmen getroffen, die Einfluss darauf

haben, wie mit Internetseiten interagiert werden kann, welche Produkte zu welchem Preis angeboten werden oder eben verborgen bleiben. Im Falle von Werbung kann man das verschmerzen oder gar positiv sehen. Wenn dieselben Mechanismen aber auch verwendet werden, um Suchmaschinenergebnisse oder sogar Nachrichten ungefragt und/oder nicht-transparent zu personalisieren, wie es Pariser (2012) mit dem Begriff der *Filter-Bubble* beschreibt, ist die Art und Weise, wie die Welt wahrgenommen wird und wie Entscheidungen gefällt werden, betroffen.

Dezisionale Privatheit ist nach Rössler (2001: 146) notwendig zum Schutz von Freiheiten, die „angewiesen [sind] auf sozialen Raum, in dem sie unbehelligt gelebt werden können“. Die beschriebenen Mechanismen sind dahingehend problematisch, als sie diesen Raum einengen bzw. die jeweilige Person darüber im Unklaren lassen, wie der Raum gestaltet ist, in dem sie sich aufhalten und welche Konsequenzen ihre Handlungen haben können oder werden. Dies trifft nicht nur auf Informationsräume zu, sondern auch dann, wenn der ganz reale Wohnraum sich etwa in einem Viertel befindet, bei dem die Wohnumfeldanalyse die Kreditwürdigkeit senkt oder wie beim *Redlining* Ausschlüsse produziert.

Rössler bezieht dezisionale Privatheit vor allem auf zwischenmenschliche Beziehungen. Sie versteht darunter etwa das Verhältnis einer Tochter gegenüber den Eltern oder eines homosexuellen Paares gegenüber Verwandten, die sich gegen Einflussnahmen auf die Entscheidung für eine Lebensweise verwehren und auf ihre Privatsphäre verweisen. Wenn sich diese Beziehungen, wie in *Social Networks*, insofern verändern, als sie als Relationen betrachtet und automatisiert gemessen werden, interagiert der Dienst mit einem Beziehungsgeflecht und wird mithin selbst Akteur.[13] Gleichzeitig versucht der Dienst dabei unsichtbar zu sein und sowohl die Entscheidungsgrundlagen (als Geschäftsgeheimnisse) zu verheimlichen als auch die Praktiken, mithilfe derer Entscheidung gefällt werden, nicht öffentlich zu machen. So wird die Möglichkeit genommen, sich dieser Einflussnahme zu erwehren, weil es kein konkretes Gegenüber gibt und die Mechanismen in einer Black Box verborgen bleiben.

[13] So wird z.B. bei *Facebook* jede Kommunikation analysiert und die Sichtbarkeit von Beiträgen der einen für die andere Partei unbemerkt verhindert, indem Beiträge als unwichtig eingestuft werden, oder durch Änderungen in den Privatssphäreneinstellungen auch erst möglich gemacht. Die beschriebene Verwehrung von Einmischung wird so mindestens innerhalb der Plattform erschwert.

Im Zuge einer Ausdifferenzierung des Privatheitsbegriffs hat Vedder (2004: 467) den Begriff der *Categorial Privacy* geprägt, bei dem individuelle Privatheit für die einzelnen Individuen einer Gruppe betrachtet wird, und die sich damit von der *collective privacy* unterscheidet. Bei jener werden die Eigenschaften von Gruppen im Rahmen einer Privatheitsdiskussionen als schützenswert betrachtet, ohne die Auswirkungen auf Einzelne bzw. die Verletzung ihrer Privatsphäre berücksichtigen. Dagegen hat

> categorial privacy [...] its points in respecting and protecting the individual rather than in respecting and protecting the group to which the individual belongs. [...] [I]t draws attention to the attribution of generalized properties to members of groups, which, however, may result in the same effects as the attribution of particularized properties to individuals as such. In this respect, categorial privacy resembles stereotyping and wrongful discrimination on the basis of stereotypes. (ebd.: 469)

Vedder argumentiert also, dass auch solche Informationen einem Privatheitsschutz unterliegen sollten, die nicht eine konkrete Person betreffen, sondern eine bestimmte Gruppe, da sie die Einzelnen, die dieser Gruppe zugerechnet werden, mit den selben Folgen behaftet, als wären es personenbezogene Informationen. Hierunter kann man auch solche Daten fassen, die zwar anonymisiert, d.h. ohne konkreten Personenbezug sind, aber dennoch eine Kategorisierung und Attribuierung der Kategorien vornehmen.

4.2 Die kybernetische Hypothese

Solche begrifflichen Ausdifferenzierungen sind hilfreich, um die Problematik zeitgenössischer Datenproduktions- und -distributionsweisen wieder an aktuelle, liberale Privatheitskonzepte zurückzubinden. Dennoch ergeben sich hier eine Reihe von Schwierigkeiten, die die Annahme einer liberalen Gesellschaft selbst betreffen. Die angeführten Konzepte von Rössler und Vedder bekennen sich deutlich zu einem Bild vom Individuum, dessen Autonomie zu schützen Aufgabe liberaler Gesellschaften ist. Neben der Fokussierung auf die einzelne Person sollen im Folgenden die beschriebenen Praxen selbst stärker in den Blick genommen werden. Dazu eignet sich insbesondere die Analyse des *kybernetischen Kapitalismus* wie ihn etwa das Kollektiv *Tiqqun* (2007) entwickelt und damit den Begriff der Kybernetik wiederbelebt hat. Die kybernetische Hypothese beschreiben sie als eine politische Hypothese,

> eine neue Fabel, welche die liberale Hypothese seit dem Zweiten Weltkrieg endgültig verdrängt hat. Im Gegensatz zu jener schlägt sie vor, die biologischen, physischen und sozialen Verhaltensweisen als voll und ganz programmiert und neu programmierbar zu betrachten. (ebd.: 13)

Die kybernetische Hypothese ersetzt oder überschreibt dabei gerade diejenige politische Idee, die die Entwicklung des Autonomie- und Privatheitsbegriffs in den letzten Jahrhunderten geprägt hat. Sie entstammt der Entwicklung der Kybernetik, die ihren Anfang Mitte des 20. Jahrhunderts auf Basis der Schriften Norbert Wieners nahm und deren Grundidee es ist, dass die Welt, wie in der Systemtheorie, aus (technischen wie sozialen) Systemen besteht und jegliches System regulierbar ist. Zugrunde liegt ein Experiment Wieners, der ein automatisiertes Flugabwehrsystem entwickeln wollte, dass die zukünftige Position eines Flugzeugs aus der bisherigen Flugbahn approximieren sollte. Die Annahme war, dass das vergangene Flugverhalten das zukünftige determiniere, weshalb die zukünftige Position ständig auf Basis der aktuellen und vergangenen Positionen neu berechnet werden sollte. Diese Idee – kontinuierliches Regulieren durch Feedback – entwickelte sich unter anderem in den Macy Konferenzen (vgl. Pias 2002) bis in die 1970er zu einer universalen Denkhypothese, die in fast allen Wissenschaften übernommen wurde, wenn auch der Begriff selbst weniger verwendet wurde.

Die kybernetische Hypothese behauptet, dass „die Kontrolle über ein System durch einen optimalen Grad der Kommunikation zwischen seinen Teilen erreicht wird“ (Tiqqun 2007: 23). Erst nach der Trennung einer Information vom Gegenstand selbst lasse sich diese Kommunikation bewerkstelligen, die eine Regulierung über Feedback erlaubt und im Anschluss auf den Gegenstand zurück wirkt. Wenn also mein Klick-Verhalten auf Webseiten abgefangen und als Information von mir als Person abgespalten wird, kann diese mittels des Algorithmus genutzt werden, um wiederum mittels Werbung als Feedback mein Verhalten zu regulieren. Tiqqun (2007: 37) stellen dazu fest: „Das Internet ermöglicht es gleichzeitig, die Präferenzen des Konsumenten zu erkennen und sie durch die Werbung zu steuern“. Aber auch die Idee des – nicht nur online stattfindenden – Kreditscorings ist in der Annahme begründet, dass sich zukünftiges Verhalten auf Basis einer ständig aktualisierten Liste vergangener Aktionen berechnen ließe. Dabei fließt auch jede Aktion, die nicht so berechnet war (man könnte auch sagen ein Fehler gegenüber der Berechnung ist), wieder in zukünftige Berechnungen ein und trägt zur Optimierung des Systems bei.

Morris (2012) identifiziert auf Basis der Arbeiten von Tiqqun und Konzepten wie dem von Deleuze entwickelten ‚Dividuum‘ Mechanismen, die besonders stark auf Basis der kybernetischen Hypothese agieren. Das Dividuum ist, wie

auch in den vorgestellten Beispielen, nur seiner Attribute nach von Interesse für die Unternehmen und für staatliche Institutionen. Dadurch wird es besser durch Mittel der Biopolitik kontrollierbar, etwa über biometrische Verfahren oder die oben beschriebenen Sicherheitsverfahren sowie „less visible forms of control like scoring or facebook face recogintion“ (Morris 2012). Die konkrete Identität ist erst dann von Relevanz, wenn es darum geht, die Person im Raum der liberalen Hypothese zu unterwerfen und als natürliche Person zur Rechenschaft zu ziehen, sei es als es als mutmaßliche Terroristinnen und Terroristen[14] oder zu Abrechnungszwecken bzw. zur Vertragserfüllung.

Es ist zu diskutieren, was dieses kybernetische Verständnis für eine liberale Gesellschaftsordnung, deren Privatheitsbegriffe sich vor allem auf einen starken Personenbegriff stützen, der mit den juristisch natürlichen Personen koinzidiert, bedeutet. Es bedeutet zwar nicht, dass der Personenbegriff aufgegeben werden muss, aber zumindest, dass er porös wird. Eine Kritik an den *Data-Mining*-Systemen auf Basis dieses Personenbegriffs ist insofern schwierig, als er den Systemen sogar weitere Möglichkeiten schafft, wenn diese abseits davon agieren und Daten anonymisiert verarbeiten.

5 Handlungsstrategien zum Umgang mit den Folgen von *Data Mining*

Ausgehend von dieser Analyse lässt sich nun fragen, wie mit den Folgen der technischen Entwicklung hin zu mehr und umfangreicheren Datenanalysen und Scorings konkret umgegangen werden kann. In Anlehnung an Lessig (2006) kann man Entwicklung auf den Gebieten des Rechts, der Normen und des *Codes*, also der technischen Ebene, gleichermaßen als Einflussgrößen auf sozio-technische Prozesse betrachten. Daher sollen im Folgenden unterschiedliche Herangehensweisen an das Problem beschrieben werden.

5.1 Rechtliche Regulierung

Insbesondere Verfahren wie das Scoring sind schon länger Gegenstand der Gesetzgebung. Die EU-Datenschutzrichtlinie von 1995 enthielt ein Verbot automatisierter Einzelentscheidungen in Fällen, bei denen mit erheblichen Beeinträchtigungen oder rechtlichen Folgen zu rechnen ist. Weniger stark

[14] Im Juni 2013 wurde bekannt, dass zur Vorbereitung gezielter Tötungen durch Drohnen vor allem die terroristischen Signaturen der Zielpersonen ausschlaggebend sind. In vielen Fällen sind daher nicht Namen und konkrete terroristische Aktionen der Getöteten bekannt, sondern vielmehr Orte an denen sie sich aufgehalten und Personen die sie gekannt haben (Kaplan 2013).

reguliert sind bisher Verfahren des Profilings. Erst im Entwurf der EU-Datenschutzgrundverordnung vom Januar 2012 und vor allem in den eingebrachten Änderungsvorschlägen wird Profiling erstmals definiert als

> jede Form automatisierter Verarbeitung personenbezogener Daten, die zu dem Zweck vorgenommen wird, bestimmte personenbezogene Aspekte, die einen Bezug zu einer natürlichen Person haben, zu bewerten, zu analysieren oder insbesondere die Leistungen der betreffenden Person bei der Arbeit, ihre wirtschaftliche Situation, ihren Aufenthaltsort, ihre Gesundheit, ihre persönlichen Vorlieben, ihre Zuverlässigkeit oder ihr Verhalten vorauszusagen.

In einem eigenen Artikel (Art. 20) wird dann nicht ein Verbot formuliert, sondern es werden Erlaubnistatbestände beschrieben, die Profilingmaßnahmen im Wesentlichen auf solche Situationen beschränken, in denen diese zur Vertragserfüllung notwendig sind, eine Einwilligung vorliegt sowie keine „besonderen personenbezogenen Daten“ (ebd.) genutzt werden, wodurch Diskriminierung verhindert werden soll.

Die Vorschläge der Artikel 29 Datenschutzgruppe (Article 29 Data Protection Working Party 2013) gehen noch etwas weiter und fordern eine explizite Einwilligung, dass erhobene Daten zu Profilingmaßnahmen genutzt werden dürfen. Sie fordern Transparenz über die Zwecke eines Profilings und über die involvierte Logik sowie die Rechte für Betroffene, die Profile einzusehen, zu verändern, deren Löschung zu verlangen sowie das Recht Entscheidungen, die auf Basis von Profiling getroffen wurden, zu widersprechen.

Trotzdem bleibt die Möglichkeit anonyme und pseudonyme Daten zu verarbeiten erhalten. Es soll vor allem verhindert werden, dass sie uneingeschränkt direkt zur Wirkung kommen. Aufgrund der notwendigen Bezugnahme auf eine betroffene Person ist es nicht möglich, den impliziten Varianten des Profilings zu widersprechen, die nicht darauf abzielen, eine konkrete natürliche Person zu beschreiben, sondern abstrakte Attributgruppen.

Profiling zu verhindern, steht derzeit weder auf der Agenda staatlicher noch privatwirtschaftlicher Akteure. Insofern ist es nicht verwunderlich, dass *Big Data* unser Verständnis von Personen, Informationen und Verhalten grundsätzlich verändert.

5.2 Änderung gesellschaftlicher Normen

Die kritische, Privatsphäre privilegierende, liberale Sicht bleibt heute nicht unhinterfragt. Eine *Post-Privacy*-Bewegung kritisiert die Abwehrhaltung des gegenwärtigen Datenschutzes: Mit Kampagnen und Begriffen wie *data love*

beziehen sie sich positiv auf die Verdatung und versprechen sich Vorteile für alle, wenn eine breite Datenverarbeitung akzeptiert statt unterbunden würde.[15] Sie erkennen damit an, dass die Informationen überhaupt eine immaterielle Ressource bilden, die Mehrwert im ökonomischen Sinn generieren kann. Zudem sind einige Forderungen der Bewegung dabei durchaus produktiv auch für eine ablehnende Perspektive auf das Versprechen von *Big Data*. Die zentrale Forderung nach Transparenz richten die Verfechterinnen und Verfechter von *Post Privacy* nicht nur an die Einzelnen, sondern auch an die Datenverarbeitenden. Sie fordern Einsicht in Vorgänge, die sich bisher gesellschaftlicher Kontrolle entziehen. Die Veröffentlichung von Kategoriensystemen und den zugrunde liegenden Datensätzen ist auch eine Forderung von Verfechterinnen und Verfechtern der informationellen Selbstbestimmung.[16] Statt den Rechten der Einzelnen werden aber Forderungen der Gemeinschaft an die datenverarbeitenden Stellen formuliert und damit steht die Person nicht mehr als Akteur und Argumentationsgrundlage im Vordergrund.

5.3 Technische Ansätze

Die Informationstechnik reagiert auf Probleme, die mit (in der Regel neuen) technischen Entwicklungen zusammenhängen. Hier sind im Wesentlichen zwei Herangehensweisen zu beobachten. Auf der einen Seite entwickeln sich im Bereich der automatisierten Auswertung Techniken, die Datenanalysen und insbesondere Datenbankabfragen ermöglichen, ohne dass personenbezogene Daten dabei einsichtig sind (*privacy preserving data mining)*. Im Zuge der Entwicklung von *differential privacy* (Dwork 2011) etwa werden

[15] In Anlehnung an die feministische Kritik des Privaten aus den 1970er Jahren kritisieren sie die autonomen Subjekte der liberalen Hypothese als eine Erfindung männlich-weißer Dominanz. Zusammen mit einer gehörigen Portion Zukunftsoptimismus leiten sie daraus ab, dass das Private v.a. dazu diene, Machtverhältnisse zu stabilisieren und daher abgeschafft gehöre. Nur so könne gesellschaftliche Veränderung stattfinden – eine Forderung, bei der zeitgenössische Feministinnen und Feministen nicht unbedingt mitgehen. Siehe zu *Post-Privacy* u.a. Heller (2011) und zu den Versprechungen von *Big Data* Mayer-Schönberger (2013).

[16] Gegen die Veröffentlichung solcher Informationen spricht aus heutiger Perspektive aber auch, dass gerade das *Black-Boxing* und die Geheimhaltung der Arbeitsprozesse sowie die marktwirtschaftliche Konkurrenz zwischen den Daten hortenden und verarbeitenden Unternehmen dazu beiträgt, dass Gandies *panoptic sort* nicht Wirklichkeit wird, da einzelne Konzerne zwar großen Datensammlungen betreiben, aber eine Zusammenführung mit den Datenbanken der Konkurrenz um jeden Preis vermieden wird.

Datenbankabfragen nicht mehr direkt auf Basis von Rohdaten durchgeführt, sondern es wird in einem Zwischenschritt auf Protokollebene sichergestellt, dass keine personenbezogenen Daten – z.B. durch geschickte Aneinanderreihung verschiedener Abfragen – aus den Ergebnissen extrahiert werden können (vgl. McSherry 2010). Beim *discrimination-aware data mining* (Pedreshi et al. 2008) liegt der Fokus auf den Klassifizierungsalgorithmen. Die Autorinnen haben versucht, Diskriminierung durch Klassifizierung messbar zu machen, indem sie solche Kategorien identifizieren, die einen übermäßig großen Einfluss auf das Ergebnis haben. Ziel solcher Anpassung der statistischen Methoden ist es, diese kompatibel insbesondere mit juristischen Anforderungen zu machen.

Auf der anderen Seite wird eine Reihe von Werkzeugen entwickelt, die im Sinne einer liberalen Privatheitslogik die Eigenständigkeit und Kontrollmöglichkeiten der Einzelnen stärken sollen, indem sie z.B. ermöglichen, Werbung auszublenden oder mittels Plugins Tracking zu unterbinden. Zu nennen sind hier passive Protokollvereinbarungen wie *DoNotTrack*[17] oder Cookie Blocker wie etwa *Ghostery*[18]. Eine Reihe weiterer Tools versucht, aktiv Einfluss nehmen und die Profilingmaßnahmen zu beeinflussen. Darunter fällt z.B. die Browsererweiterung *TrackMeNot* (Howe/Nissenbaum 2009; vgl. Toubiana et al. 2011), bei der der Browser im Hintergrund zusätzliche Anfragen stellt, um das tatsächliche Such- und Surfverhalten zwischen dem automatisch erzeugten Datenrauschen zu verbergen (*obfuscation*) und so ein genaues Interessensprofiling verhindert. Einen Schritt weiter ging der mittlerweile eingestellte Dienst *Google Sharing*[19], bei dem über einen separaten Server Identifizierungscookies von Google zwischen Internetusern ausgetauscht wurden, so dass sich das Verhalten mehrerer User für den Algorithmus vermischt (*scrambling*). Anders als bei *TrackMeNot* wird dadurch nicht nur das eine Profil unscharf gemacht, sondern auch gleichzeitig die automatisierte Kategorisierung der selbst-lernenden Algorithmen gestört.

---

17 Online verfügbar unter http://www.donottrack.us (zuletzt aufgerufen am 25.06.2013).

18 Betrieben von einem Unternehmen, das sich Kritik ausgesetzt hat, da es selbst Daten sammelt und Werbetreibenden – natürlich anonymisierte – Informationen über die Anzahl der verhinderten Werbeeinblendungen verkauft, sofern die Nutzerinnen und Nutzer einwilligen (vgl. Bilton 2012).

19 Ein nicht mehr gepflegter Dienst, der es Nutzerinnen und Nutzern erlaubte, Tracking Cookies von *Google* automatisiert auszutauschen, um die unterschiedlichen Suchprofile zu vermischen. Noch online verfügbar unter https://github.com/Abine/GoogleSharing (zuletzt aufgerufen am 25.06.2013).

Erschwert werden *obfuscation* und *scrambling* durch neue Verfahren wie das Browser-Fingerprinting (vgl. Boda et al. 2012; Eckersley 2010), bei dem nicht mehr ein Cookie zur Identifizierung unterschiedlicher Nutzerinnen und Nutzer eingesetzt wird, sondern stattdessen Informationen über das verwendete Gerät (Betriebssystem und Bildschirmauflösung) sowie Browsereinstellungen (installierte Plugins und Schriftarten) genutzt werden, um diese zu unterscheiden. Mittels *Cross-Device-Tracking* wird zusätzlich versucht, anhand gemeinsamer externer IP-Adressen und ermittelter Surfprofile unterschiedliche Geräte zu einem Gesamtprofil zusammenzuführen.

6 Zusammenfassung

Die Analyse der Methoden von *Big Data* und *Profiling* muss auch die technischen Grundlagen des *Data Mining* heranziehen, um die vermeintliche Objektivität gegenüber Fragen der Privatheit bei automatischer Datenauswertung in Zweifel ziehen zu können. Eine Konfliktlinie zwischen aktuellen Konzepten von Privatheit und den technischen Möglichkeiten wird dort deutlich, wo es kaum mehr notwendig ist, personenbezogene Daten zu verarbeiten und daher nicht auf Basis liberaler Verständnisse [wirklich Plural?] von Personen agiert wird, sondern stattdessen – einer kybernetischen Hypothese folgend – Mengen von Attributen und Informationen das Denken und Handeln bestimmen. Diese Informationen werden einerseits zur Einschätzung und Verhaltensvorhersage genutzt und sind dabei offen und verdeckt diskriminierend bzw. verfestigen bestehende Verhältnisse. Andererseits werden die Daten genutzt, um die Handlungsoptionen des oder der Einzelnen einzuschränken und, wenn möglich, zu beeinflussen. Dabei wird erst bei den konkreten Folgen der Nutzung jener Daten wieder auf konkrete Personen Bezug genommen und eine Kritik an den dazwischen liegenden Schritten entkräftet. Hierdurch werden liberale Konzepte von Privatheit, die den Schutz der informationellen und dezisionalen Selbstbestimmung in den Fokus stellen, gezielt unterlaufen. Nichtsdestotrotz gibt es rechtliche und technische Ansatzpunkte, dieser Praxis zu begegnen, den Schutz von Privatheit zu stärken und Möglichkeiten, Autonomie zu erhalten. Notwendige Interventionen sollten dabei verstärkt Perspektiven entwickeln, die gemeinschaftliche Ansätze verfolgen, anstatt einseitig auf individuelle Einwilligungen und Rechte zu setzen.

Literatur

Alvarez, L. (2010): Meet Mikey, 8: U.S. Has Him on Watch List. In: *The New York Times*, 13.01.2010. Online verfügbar unter: http://www.nytimes.com/2010/01/14/nyregion/14watchlist.html, zuletzt aufgerufen am 26.06.2013.

Article 29 Data Protection Working Party (2013): *Advice Paper on Essential Elements of a Definition and a Provision on Profiling Within the EU General Data Protection Regulation.* Online verfügbar unter http://ec.europa.eu/justice/data-protection/article-29/documentation/other-document/files/2013/20130513_advice-paper-on-profiling_en.pdf, zuletzt aktualisiert am 13.05.2013, zuletzt aufgerufen am 26.06.2013.

ARTICLE 29 Data Protection Working Party (2010): Opinion 2/2010 on Online Behavioural Advertising. Online verfügbar unter http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2010/wp171_en.pdf, zuletzt aktualisiert am, zuletzt aufgerufen am 11.11.2013.

Barnett, A. (2004): CAPPS II: The Foundation of Aviation Security? In: *Risk Analysis* 24, 4, S. 909-916.

Bäumler, H. (2002): *Tätigkeitsbericht 2002 des Landesbeauftragten für den Datenschutz.* Online verfügbar unter https://www.datenschutzzentrum.de/material/tb/tb24/kap6.htm#Tz6.2.1, zuletzt aufgerufen am 26.06.2013.

Biermann, K. (2010): Behavioral Targeting: Wie vorhersagbar unser Verhalten ist. In: *Die Zeit* 15.02.2012. Online verfügbar unter http://www.zeit.de/digital/datenschutz/2010-02/noller-nugg-targeting, zuletzt aufgerufen am 26.06.2013.

Bilton, R. (2012): *Ghostery: A Web Tracking Blocker That Actually Helps the Ad Industry*. Online verfügbar unter http://venturebeat.com/2012/07/31/ghostery-a-web-tracking-blocker-that-actually-helps-the-ad-industry/, zuletzt aktualisiert am 31.07.2012, zuletzt aufgerufen am 26.06.2013.

Boda, K. et al. (2012): User Tracking on the Web via Cross-Browser Fingerprinting. In: Laud, P. (Hg.): *Information Security Technology for Applications,* Berlin, Heidelberg, S. 31-46.

Danna, A./Gandy, O. H. (2002): All That Glitters is Not Gold: Digging Beneath the Surface of Data Mining. In: *Journal of Business Ethics* 40, 4, S. 373-386.

Deleuze, G., (1990): Postskriptum über die Kontrollgesellschaften. *L'autre journal* 1, 1, o.P.

Domingos, P. (2012): A Few Useful Things to Know About Machine Learning. In: *Communications of the ACM* 55, 10, S. 78.

Duhigg, C., (2012): How Companies Learn Your Secrets. In: *The New York Times*, 19.02.2012. Online verfügbar unter https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html, zuletzt aufgerufen am 26.06.2013.

Dwork, C. (2011): A Firm Foundation for Private Data Analysis. In: *Communications of the ACM* 54, 1, S. 86–95.

Eckersley, P. (2010): How Unique Is Your Web Browser? In: Atallah, M. J./Hopper, N. J., (Hgg.): *Privacy Enhancing Technologies*. *Lecture Notes in Computer Science*. Berlin, Heidelberg, S. 1-18.

Elias, B./Krouse, W./Rappaport, E. (2005): *Homeland Security: Air Passenger Prescreening and Counterterrorism*. Online verfügbar unter: http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA453620, zuletzt aufgerufen am 26.06.2013

Fayyad, U./Piatetsky-Shapiro, G./Smyth, P. (1996): From Data Mining to Knowledge Discovery in Databases. In: *AI Magazine* 17, 3, S. 37.

Galloway, A.R. (2004): *Protocol: How Control Exists After Decentralization*. Cambridge, MA.

Gandy, O. H. (1993): *The Panoptic Sort - A Political Economy of Personal Information*, Boulder.

Han, J./Kamber, M./Pei, J. (2011). *Data Mining: Concepts and Techniques*. Burlington, MA.

Heller, C. (2011): *Post Privacy: Prima leben ohne Privatsphäre*. München.
Howe, D.C./Nissenbaum, H. (2009): TrackMeNot: Resisting Surveillance in Web Search. In: Kerr, IanJ. (Hg.) *Lessons from the Identity Trail: Anonymity, Privacy, and Identity in a Networked Society*. Oxford, S. 417-436.
International Working Group on Data Protection in Telecommunications (2013): *Working Paper on Web Tracking and Privacy: Respect for Context, Transparency and Control Remains Essential*. Prague. Online verfügbar unter http://www.datenschutz-berlin.de/attachments/949/675.46.13.pdf, zuletzt aufgerufen am 26.06.2013.
Kamp, M./Weichert, T. (2005): *Scoringsysteme zur Beurteilung der Kreditwürdigkeit – Chancen und Risiken für Verbraucher*. Online verfügbar unter https://www.datenschutzzentrum.de/scoring/2005-studie-scoringsysteme-uld-bmvel.pdf, zuletzt aufgerufen am 26.06.2013.
Kaplan, F. (2013): Obamas rechtswidriger Krieg. In: *Technology Review*, 26.06.2013. Online verfügbar unter http://www.heise.de/tr/artikel/Obamas-rechtswidriger-Krieg-1895034.html, zuletzt aufgerufen am 26.06.2013.
Konferenz der Datenschutzbeauftragten des Bundes und der Länder (2000): *Data Warehouse, Data Mining und Datenschutz*, Online verfügbar unter http://www.datenschutz.hessen.de/k59e2.htm, zuletzt aufgerufen am 26.06.2013.
Lessig, L. (2006): *Code: And Other Laws of Cyberspace, Version 2.0*. New York. Online verfügbar unter http://www.codev2.cc/download+remix/Lessig-Codev2.pdf, zuletzt aufgerufen am 26.06.2013
Linden, G./Smith, B./York, J. (2003): Amazon.com Recommendations: Item-to-item Collaborative Filtering. In: *IEEE Internet Computing* 7, 1, S.76-80.
Mayer, J. R./Mitchell, J. C. (2012): Third-Party Web Tracking: Policy and Technology. In: *IEEE Symposium on Security and Privacy 2012*. S. 413-427.
Mayer-Schönberger, V./Cukier, K. (2013): *Big Data: A Revolution That Will Transform How We Live, Work and Think*. London.
McSherry, F. (2010): Privacy Integrated Queries. In: *Communications of the ACM* 53, 9, S. 89.
Morris, A. (2012): Whoever, Whatever: On Anonymity as Resistance to Empire. In: *Parallax* 18, 4, S. 106-120.
Oliveira, S. R./Zaïane, O. R. (2004): Toward Standardization in Privacy-preserving Data Mining. In: *Proceesdings of the 3rd Workshop on Data Mining Standards in Conjunction with KDD 2004*, Seattle, S. 7-17.
Pariser, E. (2012): *Filter Bubble: Wie wir im Internet entmündigt werden*. München.
Pedreshi, D./Ruggieri, S./Turini, F. (2008): Discrimination-aware Data Mining. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. Las Vegas, S. 560-568.
Pias, Claus (2002): Die Kybernetische Illusion. In: Liebrand, C./Schneider, I. (Hgg.): *Medien in Medien*. Köln.
Rodgers, Z. (2013): Product Manager David Baser on Facebook's Attribution Roadmap. Online verfügbar unter http://www.adexchanger.com/social-media/product-manager-david-baser-on-facebooks-attribution-roadmap/, zuletzt aktualisiert am 23.01.2013, zuletzt aufgerufen am 26.06.2013
Rössler, B. (2001): *Der Wert des Privaten*. Frankfurt am Main.
Singer, N. (2012): Acxiom, the Quiet Giant of Consumer Database Marketing. In: *The New York Times* 17.06.2012. Online verfügbar unter https://www.nytimes.com/2012/06/17/technology/acxiom-the-quiet-giant-of-consumer-database-marketing.html, zuletzt aufgerufen am 26.06.2013.
Sweeney, L. (2013): Discrimination in Online Ad Delivery. In: *Social Science Research Network*. Online verfügbar unter http://papers.ssrn.com/abstract=2208240, zuletzt aufgerufen am 26.06.2013.
Tavani, H. T. (1999): Informational Privacy, Data Mining, and the Internet. In: *Ethics and Information Technology* 1, 2, S. 137-145.
Tiqqun (2007): *Kybernetik und Revolte*, Zürich, Berlin.
Toubiana, V./Subramanian, L./Nissenbaum, H. (2011): TrackMeNot: Enhancing the Privacy of Web Search. Online verfügbar unter http://arxiv.org/abs/1109.4677, zuletzt aufgerufen am 26.06.2013.

Vedder, A. H. (2004): KDD, Privacy, Individuality, and Fairness. In: Spinello, R. A./Tavani, H. T./Vedder, A. H. (Hgg.): *Readings in Cyberethics*. Sudbury, MA.
Weichert, T. (2006): Kredit-Scoring und Datenschutz. Online verfügbar unter https://www.datenschutzzentrum.de/scoring/060404-kreditscoring.htm, zuletzt aufgerufen am 26.06.2013.