Prädiktive Privatheit: Warum wir alle „etwas zu verbergen haben“

Rainer Mühlhoff <muehlhoff@tu-berlin.de>

Pre-print Manuskript, Version: 2020-10-17

Im Erscheinen bei der Berlin-Brandenburgischen Akademie der Wissenschaften

Künstliche Intelligenz (KI) ist seit einigen Jahren wieder ein „gehypetes Thema“. Bestes Indiz dafür ist die öffentliche Aufmerksamkeit, die sich auf KI als Potenzial und Zukunftsprognose richtet. Selbstfahrende Autos, humanoide Roboter oder Betriebssysteme, in die man sich verlieben kann, rufen gleichermaßen utopische und dystopische Phantasien auf den Plan, die fließend Science Fiction übergehen. Auch ethische Zweifel werden dabei laut, die zum Beispiel fragen, in welcher Weise den autonomen Maschinen moralische Verantwortung für ihr Handeln und ein eigener moralischer Status zugesprochen werden müssen.

Weniger im Zentrum der populären KI-Narrative stehen jene schon jetzt verwendeten, datenbasierten KIs, die in die alltäglichen digitalen Medien eingewobenen sind: Suchmaschinen; Nachrichtenkuratierung in sozialen Medien; Scoring und Ranking von Menschen im Kontext von Versicherungen und Finanzdienstleistungen; psychologisches Profiling anhand von Verhaltensdaten im Marketing, am Arbeitsplatz, im Bildungsbereich, in den sozialen Netzwerken; KI-unterstützte Polizei- oder Justizarbeit. In diesen Anwendungsfeldern tritt KI als *Prognosesysteme* auf Grundlage von Machine Learning in Erscheinung. Auch als „prädiktive Analytik“ bezeichnet, bilden solche Prognosesysteme die Grundalge für algorithmisches Entscheiden, Profilbildung und soziale Selektion von Menschen.

Prognosesysteme präsentieren sich allerdings nicht als Roboter, sie bilden kein verkörpertes, sprechendes oder handelndes Gegenüber der Menschen in konkreten Interaktionssituationen. Prognosesysteme leben vielmehr in Rechenzentren, entziehen sich der Sichtbarkeit, und haben dennoch schon jetzt enorme Auswirkungen auf unser Denken, Fühlen und Handeln, in beinahe allen Bereichen der Politik, der Arbeit, des Konsums, der sozialen Beziehungen. In diesem Essay gebe ich einen kurzen Überblick, worum es sich dabei genau handelt, worin die Gefahren bestehen, und wieso demokratische Gesellschaften mit einem neuen Verständnis von Datenschutz und Privatheit darauf reagieren sollten.

KI-basierte Prognosesysteme

Mit „Prognosesystem“ beziehe ich mich auf *Machine Learning*-Modelle, die als Input eine Reihe verfügbarer Daten über ein Individuum (oder einen „Fall“) erhalten und als Ausgabe die Schätzung einer Zielvariable für dieses Individuum zurückgeben. Die Inputdaten sind dabei typischerweise große Mengen unstrukturierter Hilfsdaten, die leicht zugänglich sind,

zum Beispiel Trackingdaten (der Browser-Verlauf, Standort-Verlauf) oder Social Media Daten (Likes, Postings, Freund*innen, Gruppenmitgliedschaften); während es sich bei der Zielvariable typischerweise um schwer zugängliche oder besonders sensible Daten handelt, zum Beispiel die Bonität der betroffenen Person, Krankheiten, Suchtverhalten, politische Affinitäten, Geschlecht, sexuelle Orientierung, psychologische und emotionale Dispositionen.

In der prädiktiven Analytik möchte man also anhand leicht zugänglicher Daten schwer zugängliche Daten über Individuen abschätzen. Prädiktive Analytik entsteht überall dort, wo durch alltäglich verwendete digitale Medien massenweise Verhaltens- und Nutzungsdaten anfallen. Das liegt daran, dass prädiktive Modelle den einzelnen Fall anhand von „pattern matching“ mit Millionen anderer Fälle abgleichen, ihn einer algorithmisch bestimmten Gruppe besonders ähnlicher Fälle zuordnen und daraus eine Schätzung der unbekannten Zielvariable ableiten. In den meisten Fällen werden solche Modelle mit Verfahren des „überwachten Lernens“ trainiert: Dazu wird eine große Menge sogenannter „Trainingsdaten“ benötigt – das ist ein Datensatz, in dem für eine Kohorte von Individuen beide Datenfelder, also sowohl die Hilfsdaten als auch die sensiblen Zieldaten, erfasst sind. Solche Datensätze fallen regelmäßig im Kontext sozialer Alltagsmedien an, zum Beispiel produziert die Teilmenge aller Facebook-Nutzer*innen, die in ihrem Profil explizit Angaben über ihre sexuelle Orientierung machen, einen Trainingsdatensatz zur Abschätzung der sexuellen Identität; und die Gruppe der Individuen, von denen man gleichzeitig Zugriff auf ihren Browser-Verlauf und die Daten einer Gesundheits-App hat, produzieren Trainingsdaten für eine KI, die anhand von Browserverläufen Krankheitsdispositionen abzuschätzen lernen kann.

Sobald also *eine Gruppe von einigen Tausend Individuen* freiwillig oder unwissentlich zugleich Hilfsdaten und sensible Daten preisgibt, kann ein *Machine Learning*-Modell trainiert werden, welches Korrelationen zwischen den Hilfsdaten und den sensiblen Daten ermittelt.[1] Solche Modelle werden dann anschließend dazu verwendet, die sensible Zielvariable auch für Individuen abzuschätzen, über die nur die Hilfsdaten bekannt sind und die selbst niemals auch die sensiblen Daten über sich preisgeben würden.

Mediziner*nnen von der University of Pennsylvania haben gezeigt, dass sich mit dieser Vorgehensweise anhand von Facebook-Daten vorhersagen lässt, ob eine Nutzer*in an Krankheiten wie Depression, Psychosen, Diabetes oder Bluthochdruck leidet.[2] Facebook selbst hat bekannt gegeben, mittels künstlicher Intelligenz suizidale NutzerInnen anhand

1 Ob solche Modelle statistisch valide sind, ist hiermit nicht ausgesagt; in vielen Fällen sind sie es nicht, zum Beispiel weil die Trainingsdaten nicht repräsentativ für die relevante Grundgesamtheit sind. Entscheidend ist an dieser Stelle, dass diese Verfahren häufig ungeachtet statistischer Erwägungen trotzdem angewendet werden; siehe Mühlhoff, R. „Automatisierte Ungleichheit: Ethik der Künstlichen Intelligenz in der biopolitische Wende des Digitalen Kapitalismus“. *Deutsche Zeitschrift für Philosophie*, Nr. 6 (2020, im Erscheinen).

2 Merchant, R. M. et al. „Evaluating the Predictability of Medical Conditions from Social Media Posts“. *PLOS ONE* 14, Nr. 6 (2019).

ihrer Postings erkennen zu können. Eine weitere Studie zeigt, dass die Daten über Facebook-Likes dazu verwendet werden können, „eine Reihe höchst sensibler persönlicher Attribute vorherzusagen, darunter sexuelle Orientierung, Ethnie, religiöse und politische Ansichten, Persönlichkeitseigenschaften, Intelligenz, happiness, Suchtverhalten, Trennung der Eltern, Alter und Geschlecht“.[3]

Solche prädiktiven Analysen stoßen zum Beispiel bei Versicherungskonzernen auf großes Interesse, weil sie eine individuelle Risikobemessung erlauben. Krankenversicherungen können ihre Kunden mit Rabatten zum Gebrauch eines Fitness-Trackers motivieren, dessen Daten zentral gespeichert werden und so mit den Behandlungsdaten der Krankenkassen korreliert werden können, um individuelle Risikoprofile zu bestimmen. So genannte „Pay As You Drive“-Tarife von KFZ-Versicherungen verwenden Positions-Tracking und Beschleunigungssensoren in den Fahrzeugen, um mittels prädiktiver Analytik individuelle Versicherungsprämien in Abhängigkeit vom Fahrstil und Aufenthaltsorten zu bestimmen. Im Human-Resource-Management ist das KI-basierte Vorsortieren von Bewerber*innen bei Jobausschreibungen mittels sogenannter „hiring algorithms“ ein großes Thema; in den USA werden solche Systeme bereits heute bei einer Mehrheit der Einstellungsverfahren verwendet.[4]

Zu den ersten Anwendungen prädiktiver Analytik gehört die gezielte Werbung. So ist es einer US-amerikanischen Supermarktkette 2011 gelungen, anhand der Einkaufsdaten, die über Rabattprogramme (customer loyalty cards) gesammelt werden, schwangere Kundinnen zu identifizieren.[5] Im Credit Scoring wird ebenfalls schon lange auf prädiktive Modelle zurückgegriffen, die durch *Machine Learning* eine weitere Verfeinerung erhalten haben. „All data is credit data“ lautet der Leitspruch jener Sparte der Finanzindustrie, die mit alternativen Kreditrisikomodellen auf Grundlage von Verhaltens- und Nutzungsdaten auch noch diejenigen mit Krediten versorgen möchte, die nach klassischem Ermessen nicht kreditwürdig sind: So genannte „payday lending“-Anbieter wie das vom Ex-Google-Mitarbeiter Douglas Merrill gegründete Fintec-Unternehmen ZestFinance oder die deutsche Firma Kreditec nutzen mit KI-basierten Kreditangeboten die Ausweglosigkeit der Ärmsten der Armen aus – mit Jahreszinssätzen, die nicht selten 300% übersteigen.[6]

Biopolitik: Von der Kundenverwaltung zum Bevölkerungsmanagement

Auch wenn die Finanz- und Versicherungsbranche und das individualisierte Marketing der sozialen Medien die Entwicklung prädiktiver Analytik hauptsächlich angetrieben haben,

3 Kosinski, M. et al. „Private Traits and Attributes Are Predictable from Digital Records of Human Behavior“. *Proceedings of the National Academy of Sciences* 110, Nr. 15 (2013).

4 O’Neil, C. *Weapons of Math Destruction*. New York: Crown, 2016: S. 108, 148.

5 Duhigg, Charles. „How Companies Learn Your Secrets“. *The New York Times*, 16. Februar 2012. https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

6 Lippert, J. „ZestFinance Issues Small, High-Rate Loans, Uses Big Data to Weed out Deadbeats“. *Washington Post*, 11. Oktober 2014.

kommt es für eine aktuelle ethische und politische Besprechung dieses Themas darauf an, dass das Anwendungsfeld dieser Technologie allmählich den Bereich der B2C-Operationen überschreitet und zunehmend die Relationen zwischen Staat und Staatsbürger*innen („Government to Citizen", G2C) erfasst. Profilbildung, Risiko-Scoring und automatisierte Entscheidungsfindung werden nicht mehr nur für selektive Information, differenzielle Preisgestaltung und andere Strategien eingesetzt, die lediglich das *wirtschaftliche Verhalten von Marktteilnehmern* beeinflussen. Vielmehr werden diese Techniken nun dafür verwendet, die Beziehung des Einzelnen zum Staat, einschließlich der wohlfahrtsstaatlichen Institutionen, der Bildungs- und Gesundheitseinrichtungen, des Sicherheitsapparats und der Politik zu prägen. Das zeigt sich etwa im Feld der prädiktiven Polizeiarbeit und Kriminologie[7], in der Gesundheitsversorgung, im Bildungsbereich, im Jugendschutz[8].

Die Bereichserweiterung der KI von den B2C- zu den G2C-Relationen vollzieht sich dabei *nicht,* wie man meinen könnte, indem KI-Technologie vom privaten in den öffentlichen Sektor übertragen wird. Zu beobachten ist vielmehr eine tendenzielle *Integration staatlicher Institutionen in private Plattformen und Datennetzwerke,* denn *Machine Learning*-Technologie ist so tief mit privaten Daten-Infrastrukturen verstrickt, dass sie nicht anderswo und getrennt von den Wirtschaftsakteuren repliziert werden kann. Die Verwendung von KI im G2C-Bereich bedeutet deshalb eine Privatisierung öffentlicher Fürsorgeinfrastruktur; staatliche Institutionen und G2C-Kommunikation werden schrittweise in die Wertschöpfungslogiken des digitalen Kapitalismus integriert.

Diese Entwicklung bezeichne ich als „Biopolitik der KI".[9] Damit ist ein Staats- und Wirtschaftswesen des algorithmischen Bevölkerungsmanagements gemeint, das tief in die biologischen, ökonomischen und sozialen Prozesse integriert ist und sich als Kontrollapparat manifestiert. Die Biopolitik der KI verfährt dabei nach dem Prinzip „Gruppenhaft", das dem „pattern matching" inhärent ist und eine *prognostische* Verwaltung großer Kohorten und Menschenmengen im Stil des *Risiko-Cotrollings* ermöglicht*:* Jedes Individuum wird *individuell* (als Risikofaktor) erfasst, angesprochen und behandelt, dabei aber doch nicht aus dem Glaskäfig einer virtuellen Vergleichsgruppe entlassen.

Ein Effekt der KI-basierten algorithmischen Bevölkerungsverwaltung ist deshalb die Stabilisierung oder sogar weitere Anfachung von sozialer Ungleichheit. Virginia Eubanks hat dafür den passenden Begriff der „Automatisierung von Ungleichheit" ins Spiel gebracht:[10] durch den Einsatz der Algorithmen entsteht – teils entlang neuer, teils entlang bestehender Strukturen – ein unbemerktes soziales Gefälle. Individuen werden in Bezug auf ihren Zugriff auf staatliche Ressourcen, die Belegung mit Beschränkungen (Sicherheit, Polizeiarbeit, Jugendschutz, öffentliche Gesundheit) und ökonomische Chancen unterschiedlich be-

7 Hao, K. „AI Is Sending People to Jail—and Getting It Wrong". *MIT Technology Review*, 21. Januar 2019.

8 Eubanks, V. *Automating inequality*. New York: St. Martin's Press, 2017.

9 Mühlhoff, R. „Automatisierte Ungleichheit", a. a. O.

10 Ebd.

handelt. Damit kommt es zu einer computerisierten Form der *sozialen Selektion*, die die Gesellschaft in unsichtbare soziale Klassen unterteilt, z.B. in solche Menschen, die mutmaßlich ein Sicherheits- oder Gesundheitsrisiko darstellen, besseren oder schlechteren Zugang zu medizinischer Versorgung erhalten, aufgrund ihres Lernverhaltens in der Schule oder an der Universität privilegierten Zugang zu bestimmten Berufen erhalten, oder eher Opfer häuslicher Gewalt werden und deshalb präventiv vom Jugendschutz überwacht werden sollten.[11]

Prädiktive Privatheit

Angesichts der skizzierten Entwicklung, in der KI in den Bereich der sozialen Beziehungen, der Politik, des Sicherheits- und Justizwesens und der öffentlichen Verwaltung vordringt, stehen unsere Gesellschaften vor einer grundlegend neuen Herausforderung des Datenschutzes. Oben hat sich gezeigt, dass mittels prädiktiver Analytik anhand von leicht zugänglichen und oft anonym erhobenen Hilfsdaten (z. B. Trackingdaten, Bewegungsprofile, Social Media Daten) sensible Informationen über beliebige Individuen abgeleitet werden können – auch ohne das Wissen oder die Zustimmung dieser Individuen. In dieser Konstellation tritt nun ein neuer Typus der Verletzung von Privatsphäre zutage, mit dem viele Menschen im Alltag noch gar nicht rechnen: Klassischerweise stellt man sich Privatsphäreverletzungen als intrusiven Akt vor, in dem sensible Daten gezielt entwendet oder zweckentfremdet werden. Durch prädiktive Analytik kann die Privatsphäre eines Individuums aber verletzt werden, indem sensible Informationen aus anderen Daten *abgeleitet* oder *vorhergesagt* werden. Wir stehen deshalb angesichts prädiktiver Analytik vor einem neuen ethischen und politischen Problem: Sensible Informationen werden hier nicht durch ein Datenleck oder unerlaubte Weitergabe zuvor erhobener Daten preisgegeben, sondern durch Abschätzung von Verhaltensähnlichkeiten in einem kollektiv produzierten Datenpool.

Um den möglichen Missbrauch dieser Technologie zu problematisieren, benötigen wir ein erweitertes Verständnis von Privatsphäre, das sich auch auf *abgeschätzte,* nicht nur auf explizit *erhobene* Informationen erstreckt. Die Privatheit einer Person ist auch dann verletzt, wenn sensible Informationen ohne ihr Wissen und gegen ihren Willen durch Vergleich mit vielen anderen Personen abgeschätzt werden. Ich bezeichne diese Herangehensweise an Datenschutz als „prädiktive Privatheit“.[12]

Dieser Begriff kann den Ausgangspunkt für eine effizientere gesetzliche Regulierung von Big Data- und KI-Anwendungen bilden und tatsächlich sogar helfen, eine Lücke zu schließen, die durch bestehende Datenschutzgesetzgebung teilweise mit produziert wird: Während die Verarbeitung von Daten über geschützte Attribute wie Geschlecht, sexuelle Orientierung, Religionszugehörigkeit, Ethnie etc. in den meisten Datenschutzgesetzgebungen

11 Ebd., 127ff.

12 Mühlhoff, R. 2020. „Predictive Privacy: Towards an Applied Ethics of Data Analytics“. SSRN pre-print, https://ssrn.com/abstract=3724185.

streng geschützt ist und deshalb mit hohen rechtlichen Risiken und operativen Herausforderungen (z. B. Inforamtions- und Einwilligungsverfahren, sichere Datenspeicherung, …) verbunden ist, gehen Unternehmen dazu über, statt dieser Datenfelder mit sogenannten „proxies“ zu arbeiten. Dabei handelt es sich um algorithmisch bestimmte Kombinationen von (ungeschützten) Hilfsdaten, die mittels prädiktiver Analytik eine Vorhersage über die geschützten Attribute zulassen. Prädiktive Analytik birgt also nicht nur neue, „speziellere“ Möglichkeiten der Privatsphäreverletzung, sondern kann auch dafür eingesetzt werden, das bestehende Datenschutzniveau zu unterlaufen.

Um den Schutz prädiktiver Privatheit wirkungsvoll in Regulierung umsetzen zu können, benötigen wir zuerst ein breites gesellschaftliches Bewusstsein für den Mechanismus prädiktiver Privatsphäreverletzungen. Dazu gehört ein Bewusstsein dafür, dass KI-basierte Prognosen nur möglich sind, wenn und weil viele Bürger*innen sowie politische Entscheidungsträger*innen kein Bedenken darin sehen, ihre Daten freiwillig (und ggfs. anonymisiert) zur Verfügung zu stellen. „Ich habe doch nichts zu verbergen“ ist eine weit verbreitete moralische Haltung, die es großen Plattformunternehmen überhaupt erst ermöglicht, umfassende Trainingsdatensätze zu generieren. Wie oben ausgeführt, wird zum Training eines prädiktiven Modells eine Gruppe von Nutzer*innen benötigt, die sowohl die Hilfsdaten *als auch* die sensiblen Zieldaten über sich preisgeben. Auf gesellschaftlichem Maßstab betrachtet reicht dafür in vielen Fällen eine Minderheit aus, die für sich keine Probleme bei der Preisgabe dieser Daten sieht oder sich ihrer nicht bewusst ist. Oft ist diese Einstellung bei Personen zu finden, die sich selbst für „normal“ halten, nicht davon ausgehen, dass sie in den Fokus von „Überwachung“ gelangen könnten, und die selbst in ihrer gesellschaftlichen Position keine negativen Auswirkungen durch prädiktive Analytik erlebt haben. Doch nur anhand der Daten vieler „normaler“ Nutzer_innen, die meinen „nichts zu verbergen zu haben“, lassen sich die prädiktiven Algorithmen trainieren, die *andere* Individuen als Abweichler_innen erkennen können. Sich zu informieren und der Auswirkungen der des eigenen Umgangs mit sensiblen Daten für dritte bewusst zu werden, ist die ethische Dimension prädiktiver Privatheit – die uns alle betrifft. Auch die weltweit fortschrittliche europäische Datenschutzgrundverordnung ist gegen die Gefahren von Big Data und KI weitestgehend wirkungslos,[13] insbesondere, wenn sie dadurch zustande kommen, dass zahlreiche Nutzer*innen in die Verarbeitung ihrer sensiblen Daten einwilligen.

Die – weitestgehend unregulierte – Existenz KI-basierter Prognosesysteme liegt also an einer weit verbreiteten liberalistischen Haltung beim Thema Datenschutz: Jede*r kann nach eigenem Ermessen entscheiden, was er mir seinen Daten macht. Diese Herangehensweise ist eng mit der dem klassischen bürgerlichen Verständnis von Privatheit verknüpft, das auf die Kontrolle des Zugangs zur eigenen privaten Sphäre zugerichtet ist. Doch im Zeitalter von Big Data und prädiktiver Analytik kommt dieses individualistische Prinzip an seine Grenzen: Daten, die man selbst freiwillig weitergibt, können dazu verwendet werden, sen-

13 Wachter, Sandra. 2019. „Data Protection in the Age of Big Data“. *Nature Electronics* 2 (1): 6–7.

sible Informationen über *andere* Menschen abzuschätzen; und umgekehrt kann man selbst aufgrund der Daten, die *andere* über sich preisgeben, unterschiedlich behandelt werden. Es kann uns also nicht egal sein, wie unsere Mitmenschen mit ihren Daten umgehen. Und weil die negativen Auswirkungen prädiktiver Analytik nicht auf alle Gesellschaftsmitglieder gleich verteilt sind, sondern überproportional die Armen, weniger Gebildeten, Schwachen, Kranken und sozioökonomisch Benachteiligten treffen, stehen demokratische Gesellschaften hier in einer *kollektiven Verantwortung:* Wir alle müssen dafür sorgen, dass mit unseren Daten kein Missbrauch getrieben werden kann – und es ist eine gute Faustregel, davon auszugehen, dass ein solcher Missbrauch in den meisten Fällen *nicht* uns selbst trifft.