Osnabrücker Universitätsreden

Rainer Mühlhoff
Die Macht der Daten

Universitätsverlag Osnabrück
V&R unipress

Osnabrücker Universitätsreden

Gedruckt mit Unterstützung der
Universitätsgesellschaft Osnabrück e.V.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Osnabrücker Universitätsreden

Band 10

Die Präsidentin
der Universität Osnabrück
(Herausgeberin)

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Rainer Mühlhoff

Die Macht der Daten

Warum künstliche Intelligenz eine Frage der Ethik ist

Universitätsverlag Osnabrück
V&R unipress

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Der vorliegende Text ist eine erweiterte Fassung der am 19. Mai 2022 in der Aula des Osnabrücker Schlosses gehaltenen Osnabrücker Universitätsrede.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Die Macht der Daten. Warum künstliche Intelligenz eine Frage der Ethik ist

Einleitung

Die Macht der Daten – der Titel des heutigen Vortrags birgt bereits seine zwei Kernthesen. Erstens beruht künstliche Intelligenz, wie wir sie heute kennen und wie sie spürbar unsere Gesellschaften prägt, wesentlich auf Daten, und zwar auf *unseren* Daten – in welcher Form genau, das werde ich gleich diskutieren. Zweitens setzt ein ethischer Umgang mit dieser Technologie und mit den zugrunde liegenden Daten ein Bewusstsein über die damit verbundenen Machtverhältnisse zwischen Unternehmen und Individuen voraus. Auch was das bedeutet, wird sich im Verlauf dieses Vortrags zeigen.

So ist es mein Ziel für heute Abend, zu zeigen, dass hinter den vermeintlich *technischen* Entwicklungen, die wir »künstliche Intelligenz« nennen, soziale Machtfragen stehen. KI in der heutigen Form wäre nicht möglich ohne umfassende gesellschaftliche Transformationen. Diese Transformationen strahlen auf zahlreiche gesellschaftliche Bereiche aus. KI bringt neue Formen des Regierens, der kapitalistischen Akkumulation, des sozialen Miteinanders, der Arbeit und der Weltbeziehungen hervor. Diese Entwicklungen, obwohl bereits in vollem Gang, sind noch nicht ganz im gesellschaftlichen Bewusstsein angekommen und fordern dringend eine ethische Auseinandersetzung.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Teil 1: Von welcher KI sprechen wir?

Unsere Reise in diesen Themenkomplex beginnt sinnvollerweise mit einer Klärung des Gegenstandsbereichs: Was ist eigentlich künstliche Intelligenz? Da künstliche Intelligenz ein jahrhunderte-, vielleicht jahrtausendealtes Thema in Wissenschaft, Philosophie, Kunst und Kultur ist, das sich mit der Entwicklung der Technik stets verändert hat, lässt sich diese Frage kaum allgemeingültig beantworten.[1] Weil jedoch im Kontext dieses Vortrags das Interesse an unserer Gegenwart, also an KI in der *heutigen* Zeit im Vordergrund steht, möchte ich die Frage nach der Begriffsbestimmung anders stellen: Von was für einer künstlichen Intelligenz ist im Folgenden die Rede? Wie sollten wir künstliche Intelligenz *heute* verstehen, um eine möglichst relevante und treffende ethische Debatte zu ermöglichen?

A. KI im öffentlichen Diskurs

Jede gute Ethik der künstlichen Intelligenz beginnt mit genau dieser Frage des vorausgesetzten Verständnisses von KI. Denn bereits in diesem Verständnis liegt ein ethisches Framing. Das liegt daran, dass wir es bei KI mit einem unscharfen Begriff zu tun haben, der sich gleichermaßen auf technologische wie fiktionale Schöpfungen, auf tatsächliche Errungenschaften mit realen gesellschaftlichen Folgen und auf Zukunftsvisionen beziehen kann. Charakteristisch für den weitläufigen Begriff der KI ist, dass ihm etwas Verheißungsvolles, in die Zukunft Weisendes anhaftet und dass dabei schwer zu

1 Vgl. Cave u. a. 2020.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

unterscheiden ist, welche Vorstellungen, die in Nachrichten, politischen Programmen, Werbung oder Unternehmenskommunikation mit KI verbunden werden, realistisch oder übertrieben, bereits einlösbar oder bloße Wunschvorstellungen sind. Wie sich in diesem Vortrag zeigen wird, betreffen viele der ethisch brisantesten Probleme von KI-Technologie gerade jene Anwendungen von KI, die im öffentlichen Diskurs *nicht* oder nur wenig präsent sind – die also gewissermaßen im blinden Fleck unseres kulturellen Bewusstseins von KI operieren.

Populäre Beispiele für künstliche Intelligenz werden immer wieder in Feuilletons, Wissenschaftsjournalismus oder Marketing präsentiert. Zu solchen Gemeinplätzen gehören selbstfahrende Autos, Pflegeroboter, Polizeiroboter, Chatbots oder KIs, die Brettspiele spielen können, wie zum Beispiel die Alpha-Go-KI, die 2016 den Go-Weltmeister Lee Sedol geschlagen hat. Bei selbstfahrenden Autos hört man oft von dem ethischen Dilemma des sogenannten »Trolley-Problems«. Das Trolley-Problem ist ein Entscheidungsdilemma, bei dem sich ein autonomer KI-Apparat in der fiktiven Situation eines ausweglos bevorstehenden tödlichen Unfalls befindet, jedoch durch seine Entscheidungen noch beeinflussen kann, welche der verschiedenen in die Situation involvierten Personen oder Personengruppen getötet wird.[2] Auch Pflegeroboter werden häufig unter der Stichwortkombination von »Ethik« und »KI« diskutiert:[3] Ist es gut oder schlecht, für die Pflege alter Menschen Roboter zu verwenden? Fördert oder schmälert das ein würdiges Altern? Kann man »Pflege« und »Fürsorge« überhaupt an Roboter auslagern

2 Vgl. kritisch: Matzner 2019.
3 Siehe zum Beispiel Misselhorn 2018.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

oder sind sie wesentlich an die menschliche Intersubjektivität geknüpft? – Das alles sind Themen, die aktuell medial sehr stark mit dem Stichwort Ethik der KI assoziiert werden.

Aus verschiedenen Gründen wird es um diese Probleme in diesem Vortrag *nicht* gehen. Und zwar weil ich die These vertrete, dass diese Themen gemessen an ihrer relativen Dringlichkeit ein Ablenkungsmanöver gegenüber aktuell noch viel gravierenderen Auswirkungen und deshalb ethisch drängenderen Problemen von KI-Technologie darstellen. Damit meine ich Anwendungen von KI-Technologie, die schon jetzt in unseren Gesellschaften präsent sind, mit denen schon jetzt Million Menschen weltweit und in jeder Minute konfrontiert sind – und die dennoch unserer Aufmerksamkeit noch viel zu sehr entgehen. Es handelt sich dabei um datenbasierte KI-Produkte, die in vernetzte digitale Medien und Dienste eingebunden sind, die wir alle an unseren Smartphones und Computern täglich benutzen oder denen wir ausgesetzt sind: Suchmaschinen, Newsfeeds auf Social Media, Sprach- und Gesichtserkennung, Scoring-Algorithmen für Kredit- und Versicherungsrisiken, individualisierte Werbung, KI-Algorithmen im Human Ressource Management, bei Job-Auswahlverfahren und in der Personalführung, KI-Algorithmen im Sicherheits- und Polizeiapparat (*predictive policing*), im Bildungssektor, in der Medizin, im Jugendschutz.

Grundsätzlich hat die öffentliche und populäre Darstellung von KI-Technologie einen starken Hang zu KI als Zukunftsmusik und Zukunftsvision: Es werden bevorzugt Technologien besprochen, die es noch nicht oder nicht in großer Verbreitung gibt, die man sich aber vorstellt – sei es utopisch oder dystopisch gefärbt. Pfle-

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

geroboter sind Zukunftsmusik, selbstfahrende Autos als etwas, das in Massen unsere Straßen bevölkert, sind Zukunftsmusik.

Des Weiteren stehen in der populären Darstellung von KI-Technologien meistens solche Artefakte im Vordergrund, die in irgendeiner Form *verkörpert* sind – die also als autonome Entitäten, als Auto, als Roboter, gegebenenfalls auch als virtueller Chatbot uns Menschen handelnd *gegenüberstehen*. Es ist meist die Rede von KI-Systemen, die als neue Entitäten unsere Welt bevölkern werden, die auf Augenhöhe mit uns interagieren und im Extremfall sogar zu Mitgliedern unserer Gesellschaft werden. Wir haben im populären Diskurs also die Tendenz, die künstliche Intelligenz im *Inneren* autonomer technischer Agenten zu lokalisieren, die uns handelnd gegenübertreten, wie man es gut bei selbstfahrenden Autos und Pflegerobotern sieht.

Diese beiden Tendenzen, das zeitliche Framing von KI als Zukunftstechnologie und das räumliche Framing von KI als in autonomen Agenten lokalisierbare Intelligenzleistung, werden dem Status quo der KI-Technologie überhaupt nicht gerecht. Es gibt sehr viele aktuell verbreitete, bereits wirkungsmächtige Anwendungen von KI-Technologie, die diesen Vorannahmen nicht entsprechen und deshalb in gesellschaftlichen und ethischen Debatten leicht unbeachtet bleiben, obwohl sie einen enormen Einfluss auf viele von uns haben und deshalb dringend ethisch besprechenswert wären. Die meiste KI-Technologie heute wird in Rechenzentren und sozialen Medien betrieben, ohne uns als Roboter gegenüberzutreten. Solche Techniken beeinflussen zum Beispiel, welche Nachrichten, Werbeanzeigen oder Postings von unseren Freund:innen wir sehen, welche Konditionen wir

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

bei Bankgeschäften oder Versicherungen angeboten bekommen oder ob wir zu einem Job-Interview eingeladen werden. Die Artefakte, die diese Entscheidungen treffen, bekommen wir nicht als materielle, körperliche Agenten zu fassen. Es handelt sich bei diesen KI-Systemen nicht um Interaktionspartner, sondern um strukturelle Faktoren unserer digitalen Kommunikation.[4] Sie beeinflussen unseren Zugriff auf Informationen, Ressourcen und Chancen, indem sie dafür eingesetzt werden, über unser Verhalten, unsere Gedanken, Gefühle, Krankheiten, Wünsche und Probleme Vorhersagen zu treffen, die es den Systemen erlauben, jede:n von uns individualisiert – und das heißt unterschiedlich – zu behandeln.

Um der aktuellen Realität von KI-Technologie gerecht zu werden, benötigen wir ein anderes Verständnis von KI. Es darf KI nicht verkörpern und nicht als Zukunftsvision framen, sondern muss die bereits jetzt aktuelle, komplexe Verwobenheit von Menschen und KI-Systemen jenseits klassischer Interaktionsbeziehungen sichtbar machen. Ich möchte Ihnen in diesem ersten Abschnitt anhand einiger Beispiele erläutern, was ich mit dieser komplexen Verwobenheit meine, und darauf aufbauend ein Verständnis von KI-Systemen als vernetzte Mensch-Maschine-Systeme, als *Human-Aided AIs*, wie ich es nenne, einführen.

4 So ausführlicher argumentiert in Mühlhoff 2020b.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Suchen Sie einmal mit der Google-Bildersuche[5] nach einem Stichwort Ihrer Wahl – zum Beispiel nach »rosa Elefant«. In dem Moment, wo Sie die Ergebnisse dieser Suche angezeigt bekommen, sehen Sie ein Resultat künstlicher Intelligenz – denn ein KI-System ermittelt die Bilder, die zu Ihrer Suche passen (vgl. Abbildung 1). Das ist zunächst ein Beispiel für ein KI-System, das jetzt schon da ist, also nicht in der Zukunft stattfindet; es ist sogar ein sehr wertvolles KI-System, das eine enorme Marktposition für sich behauptet. Außerdem ist es ein Beispiel für eine KI, die Ihnen nicht als körperliche, interaktive Entität entgegentritt, sondern immateriell und räumlich nicht lokalisierbar irgendwie in den Rechenzentren von Google betrieben zu werden scheint.

Wie wird ein Computersystem eigentlich dazu in die Lage versetzt, Bilder zum Stichwort »rosa Elefant« oder zu beliebigen anderen Stichworten, die Sie oben in die Textbox schreiben können, auszugeben? Das Problem der maschinellen Bilderkennung, also etwa der Wiedergabe des Inhalts eines Bildes in Wörtern, war für lange Zeit eines der hartnäckigsten Probleme im Bereich der maschinellen Intelligenz.

Um zu beschreiben, wie aktuelle KI-Technologie diese Nuss geknackt hat und wie eine solche Bildersuche zu ihrer »Intelligenz« gekommen ist, lohnt sich ein keiner Ausflug zurück an den Anfang der 2000er Jahre. Sie erinnern sich, da gab es noch keine Smartphones und kein Facebook. Die Dotcom-Blase war gerade geplatzt und

5 https://images.google.de.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

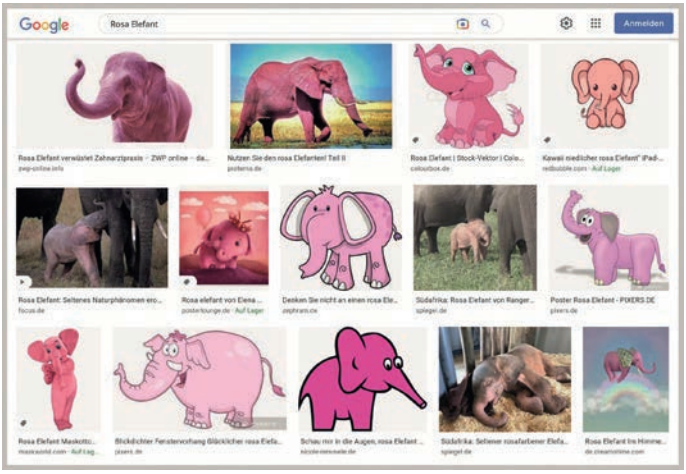


Abb. 1: Google-Bildersuche nach »rosa Elephant«, Screenshot des Autors, www.google.de, 7.10.2022.

eine Pleitewelle internetbasierter Geschäftsmodelle und Firmen griff um sich. Die Unternehmen, die aus dem Ende der ersten Internet-Ära als Überlebende hervorgingen, das sind heute große Häuser wie Amazon oder Google, die in den 2000er Jahren ihre Hauptwachstumsphase hatten. Auch Facebook kam Mitte der 2000er Jahre dazu. Von dieser Zeit möchte ich sprechen.

Damals hat Luis von Ahn, zu dem Zeitpunkt Doktorand an der Carnegie Mellon University in Pittsburgh in den USA, also an einer renommierten IT-Universität, eine Dissertation geschrieben mit dem bemerkenswerten Titel »Human computation«.[6] Dieser Begriff lässt sich kaum treffend ins Deutsche übersetzen: »menschliches Rechnen«, »menschliche Berechnung«, »Rechnen von und mit Menschen«, »menschliche Rechenwerke« – wie auch immer, ich verwende fortan »Human computation«.

6 von Ahn 2005; siehe auch von Ahn 2006a.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Luis von Ahn hat im Jahr 2006 auf Einladung von Google einen Vortrag über seine Dissertation gehalten, der auch heute noch äußerst hörenswert ist.[7] In diesem Vortrag beschreibt er retrospektiv die Idee seines Dissertationsprojektes mit den Worten: »Wir werden in diesem Projekt die gesamte Menschheit als eine extrem fortgeschrittene, großskalige, distribuierte Prozessoreinheit betrachten, die Probleme lösen kann, die Computer aktuell noch nicht lösen können.«[8] Also die gesamte Menschheit betrachtet Doktorand von Ahn in seinem Projekt als ein »großes Rechennetzwerk«. – Was für eine Idee!

Gamification: Das ESP-Game

Als Teil seines Dissertationsprojektes hat Luis von Ahn zum Beispiel ein Online-Spiel entwickelt, das unter dem Namen »ESP-Game« publiziert wurde.[9] Das war ein Browser-Spiel, das ein wenig an eine Kombination aus den bekannten Party-Spielen Tabu und Montagsmaler erinnert. Im ESP-Game spielen Sie mit einer zufällig vom Server zugeordneten anderen Person zusammen, die Sie aber nicht kennen und mit der Sie nicht kommunizieren können (vermutlich ist diese andere Person ganz woanders auf der Welt). Das Spiel besteht aus mehreren Wiederholungen derselben Aufgabe: Ihnen und Ihrer unbekannten Spielpartner:in wird auf Ihren Computer-

7 von Ahn 2006b.

8 Wortlaut im Original: »We are going to consider all of humanity as an extremely advanced, large-scale distributed processing unit that can solve large-scale problems that computers cannot yet solve.« von Ahn 2006b: 8 min.

9 von Ahn und Dabbish 2004.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Abb. 2: Screenshot ESP-Game 2003. Quelle: von Ahn und Dabbish 2004, Figure 2.

bildschirmen ein Bild angezeigt und die Aufgabe besteht darin, dass Sie beide ein Stichwort finden und eintippen müssen, welches das Bild beschreibt (siehe Abbildung 2). Wenn Sie beide exakt das gleiche Stichwort eingeben, dann bekommen Sie beide Punkte. Und je schneller Sie das tun, desto mehr Punkte bekommen Sie.

Durch diese Spielstruktur werden Sie motiviert, sich zu überlegen, was wohl das treffendste Stichwort für dieses Bild ist. Was ist das Bildmerkmal, das als allererstes in den Blick fällt? Worauf wird wohl die andere Person am ehesten kommen, wenn sie dieses Bild sieht? Das Spiel kann darüber hinaus bestimmte Stichworte als Tabu-wörter auflisten (siehe Abbildung), das macht dann die Sache minimal komplizierter, denn diese Stichworte dürfen Sie beide nicht eingeben. Durch die Tabuwörter werden die Spieler:innen gezwungen, darüber nachzu-

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

denken, welches das nächstbeste Stichwort wäre, welches das Bild treffend beschreibt.

Luis von Ahn und seine Kolleg:innen haben das ESP-Game im Jahr 2003 veröffentlicht. Es wurde relativ schnell populär und hatte schon nach vier Monaten fast 14.000 Nutzer:innen. Das klingt zunächst nach einer erfolgreichen Geschäftsidee für ein Online-Computerspiel, doch warum war das Teil einer Dissertation in Informatik? Wofür wurde das ESP-Game gemacht? – Der eigentliche Zweck dieses Spiels lag darin, Labels für den Bildinhalt einer großen Datenbank von Bildern zu erzeugen. Also für beliebige Bilder eine akkurate Liste deskriptiver Stichwörter zu erhalten, die den Bildinhalt wiedergeben. So haben bereits die 14.000 Nutzer:innen der ersten vier Monate knapp 1,3 Millionen Labels für ca. 290.000 Bilder »erzeugt«.[10]

Mitte der 2000er Jahre gab es noch keine Google-Bildersuche – beziehungsweise es gab eine Google-Bildersuche, aber wissen Sie, wie die damals funktioniert hat? Wenn Sie nach »Hund« gesucht haben, hat die Google-Bildersuche alle Bild-Dateien im Internet ausgegeben, die das Wort »Hund« im Dateinamen trugen (z. B. Hund.gif, mein-kleiner-hund.jpg), oder Bilddateien, die in Websites eingebunden waren, in denen das Wort »Hund« auftauchte. Der damals verwendete Suchalgorithmus konnte also gar nicht in die Bilder »hineinschauen«, sondern nur den sprachlichen Kontext des Bildes semantisch bewerten.

In dieser Zeit, in der kein Computersystem zu wirklicher Bilderkennung fähig war, hat das ESP-Game nun Millionen von Nutzer:innen dazu gebracht, völlig kosten-

10 von Ahn und Dabbish 2004.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

los und in kürzester Zeit qualitativ hochwertige Labels für initial 290.000 Bilder zu produzieren. Und diese Labels bilden den perfekten Suchindex für eine Bildersuche in diesen 290.000 Bildern. Das heißt, man kann die Labels dafür verwenden, von einem Stichwort, nach dem gesucht wird, auf die Bilder zu schließen, welche mit dem Stichwort assoziiert werden. Man kann die Relevanz der Bilder für ein bestimmtes Suchstichwort sogar gewichten, indem man zuerst jene Bilder anzeigt, bei denen das Stichwort im ESP-Game sehr früh erfasst wurde (ohne viele Tabu-Wörter). Luis von Ahn hat die Nutzer:innen nicht dafür bezahlen müssen, an diesen umfassenden Suchindex für seine Bilddatenbank zu kommen, denn die Spieler:innen haben ihn als Beiprodukt ihrer Spielaktivität erzeugt. In seinem Vortrag stellt von Ahn recht zynisch fest, dass er sich von den Spieler:innen für das Spiel sogar hätte bezahlen lassen können, so beliebt war es bei einigen.

Dieses Produkt seiner Dissertation, welches einen enormen Fortschritt für die Technologie der Bilderkennung versprach, wurde dann auch im Jahr 2006 von Google aufgekauft und unter dem neuen Namen Google Image Labeler publiziert. Das Spiel wurde schließlich von Google dazu verwendet, die gesamte Bilderdatenbank der Google-Bildersuche zu indizieren – also das KI-System der Google-Bildersuche zu trainieren. Der Google Image Labeler hatte so viele Spieler:innen, dass der enorme Bildbestand von Google – 425 Millionen Bilder im Jahr 2004 – in nur 6 Monaten indiziert werden konnte. Die »künstliche Intelligenz« dieser Bildersuche beruht also auf der kostenlosen Mitarbeit ganz vieler Menschen, die einfach, indem sie dieses Spiel gespielt haben, zu der Intelligenzleistung der Bildersuche beigetragen haben.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Diese kleine Vignette ist paradigmatisch für das, worüber ich spreche, wenn in diesem Vortrag von künstlicher Intelligenz die Rede ist. Ich beziehe mich dann auf KI-Systeme, die es heute schon gibt, die wir alle verwenden und zu denen wir potenziell sogar alle beitragen. Ich nenne diese Form von KI *Human-Aided AI* – menschengestützte KI. Dabei handelt es sich immer um datenbasierte KI-Systeme, also KI-Systeme, die unsere Datenspuren im Internet verwenden und nur durch sie überhaupt betrieben werden können.

C. Human-Aided AI – Menschengestützte KI

Human-Aided AI ist ein Forschungsparadigma für eine kritische Philosophie und Ethik der künstlichen Intelligenz.[11] Hinter dem Begriff verbirgt sich eine neue Perspektive auf datenbasierte KI-Systeme und eine neue Herangehensweise, um die Systeme auf ihre sozialen, gesellschaftlichen, ethischen und politischen Implikationen hin zu untersuchen. Kennzeichnend für diese Herangehensweise ist, KI-Systeme als soziotechnische Systeme zu betrachten, das heißt daraufhin auszuleuchten, inwiefern menschliche »Mitarbeit« ein integraler Bestandteil dieser Systeme ist. Human-Aided AI betrachtet datenbasierte KI-Systeme als hybride Netzwerke, in denen technische Komponenten, wie zum Beispiel Computerchips und vernetzte digitale Geräte, mit Menschen, also menschlichen Gehirnen, wenn Sie so wollen, verschaltet werden. Im Zusammenspiel bilden diese Komponenten ein kollaboratives Intelligenznetzwerk, welches

11 Die Darstellung in diesem Kapitel beruht auf Mühlhoff 2020a, 2019b.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

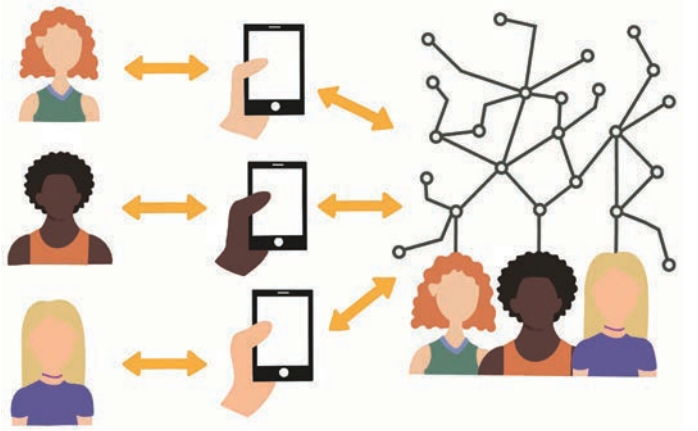


Abb. 3: Über vernetzte Interfaces werden Menschen in Human-Aided AI-Netzwerke eingebunden. Grafik: Karla Baublys.

insgesamt die Intelligenzleistung vollbringt, die wir fälschlicherweise als *rein technisch erzeugte* Intelligenzleistung wahrnehmen – etwa wenn die Google-Bildersuche uns Bilder ausgibt oder ein Übersetzungsprogramm wie DeepL oder Google Translate unsere Texte übersetzt.

Die grafischen Interfaces vernetzter Medien spielen beim Aufbau solcher hybrider soziotechnischer Systeme eine zentrale Rolle, denn über Interfaces, also Schnittstellen zwischen vernetzter Dateninfrastruktur und menschlichen Körpern, werden Menschen und ihre kognitiven Leistungen in Human-Aided AI-Netzwerke eingebunden (Abbildung 3). Die Benutzeroberflächen der Google-Bildersuche, des ESP-Games, aber auch unserer Smartphones, Tablets, Wearables und Computer etc. sind solche Interfaces. Sie sehen bereits in der oben vorgestellten Vignette, dass solche vernetzten Medien tendenziell darauf abzielen können, implizit Daten zu sammeln, die bei der Benutzung eines Service entstehen. Auf diese Weise werden die Nutzer:innen oft unbemerkt in hybride

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

KI-Systeme eingebunden – in »großskalige verteilte Rechennetze«, wie Luis von Ahn es formulierte.

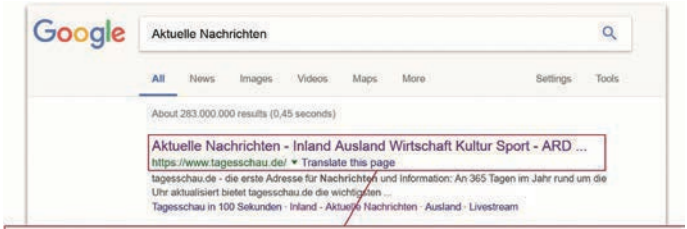
Noch einmal: Die Google-Suche

Um ein weiteres charakteristisches Merkmal von Human-Aided AI herauszustellen, lassen Sie uns noch einmal über Google sprechen, diesmal aber über die allgemeine Suchmaschine. Beim Benutzen der Google-Suche zapfen Sie ebenfalls eine KI an, die auf Suchstichworte hin eine Liste von Websites als Suchresultate produziert. Diese Suchresultate sollen aktuell und relevant sein – da trauen Sie dem KI-System ganz schön viel zu, denn woher soll eine KI wissen und beurteilen können, was gerade aktuell und relevant ist?

Die Antwort lautet: Das weiß die Google-Suche von *uns* – den Nutzer:innen. Denn wir werden bei der Benutzung der Suchmaschine zu Teilen dieses KI-Systems, das nämlich ein menschengestütztes KI-System ist. Aber wieso? Wie funktioniert das?

Wenn Sie auf eines der von Google vorgeschlagenen Suchergebnisse klicken, öffnet sich nicht nur die Seite, die Sie da angeklickt haben, sondern es passiert unbemerkt im Hintergrund noch Folgendes: Der Link, den Sie anklicken, ist so konstruiert, dass er gar nicht direkt auf die gewünschte Zieladresse führt, sondern zu einem Google-Server (Abbildung 4). Der leitet Sie dann blitzschnell an die Zieladresse weiter. Überdies enthält der Link, den Sie aufrufen, einige dynamische Variablen oder »Parameter«, wie man etwas technischer sagen würde, die auf diese Weise, ohne dass Sie es mitbekommen, von Ihrem Browser an Google übermittelt werden. Diese Parameter geben zum Beispiel an, ob es das erste Such-

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523




```
<a href="https://www.google.de/url?sa=t&ct=j&q=&esrc=s
     &source=web&cd=1&cad=rja&uact=8
     &ved=0ahUKEwjj5t3t0YLZAhULlSwKHb-eAYcQFgguMAA
     &url=http://www.tagesschau.de/
     &usg=AOvVaw1_8NtrWpEbLefVcTd78eDR">Aktuelle Nachrichten
          -Inland Ausland Wirtschaft Kultur Sport - ARD
     ...</a>
```

Abb. 4: Klicktracking-Mechanismus in der Google-Suche. Grafik: Rainer Mühlhoff / Screenshot www.google.de, 26.09.2018.

resultat ist, das Sie aufrufen, oder ob Sie sich vorher schon ein anderes angeschaut haben und dann zum Beispiel mit dem Zurückbutton wieder auf die Resultateseite gekommen sind – das wäre ja ein Indiz dafür, dass Sie mit dem ersten gesehenen Resultat nicht zufrieden waren und noch ein weiteres Ergebnis anschauen wollten.[12] Google kann ebenfalls registrieren, ob Sie im Browser die Tabs gewechselt haben zwischen verschiedenen Suchresultaten. Natürlich wird auch erfasst, wie weit Sie heruntergescrollt haben, um das angeklickte Suchresultat zu finden. Und falls Sie parallel in einem anderen Browserfenster bei Google eingeloggt sind (z. B. in Gmail, YouTube oder einem anderen Google-Service), dann wird über die mitgeschickten Parameter von Google sogar registriert, wer Sie sind – die Suche wird dann der persönlichen Suchhistorie Ihres Accounts zugeordnet.

12 Siehe dazu ausführlich: Mühlhoff 2019a.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Nachdem Google diese Parameter und noch viele weitere registriert hat, werden Sie dann ganz schnell, ohne es zu merken, auf die Zielseite weitergeleitet. Das heißt, in dem Sie die Google-Suche benutzen, zapfen Sie nicht nur ein KI-System an, sondern Sie tragen auch zu diesem KI-System bei. Denn Google lernt aus den unbemerkt erfassten Rückmeldungen, welche Resultate für bestimmte Personengruppen die relevantesten sind. Dieser enorm reichhaltige Datenbestand, der bei jedem Klick auf ein Suchresultat weltweit weiter anwächst, wird ständig in das maschinelle Lernverfahren wieder eingespeist, welches die Suchresultate hervorbringt. Das heißt, bei der Benutzung dieser KI produzieren die Nutzer:innen zugleich Trainingsdaten, die dazu verwendet werden, das System zu rekalibrieren. Nur so schafft es dieses KI-System, überhaupt aktuell zu bleiben und auch in ständig veränderten Informationskontexten (aktuelle Nachrichtensituation, politische Entwicklungen etc.) stets relevante Resultate anzuzeigen. Über das Click-Tracking, das in den Links auf der Suchresultateliste eingebaut ist, werden also Feedback-Schleifen installiert. Durch verstärkendes oder abschwächendes Feedback wird kalibriert, welche Resultate relevant sind und welche nicht (mehr). Diese Feedback-Schleifen machen die Nutzer:innen unwissentlich zu Mitwirkenden an der Intelligenzleistung des KI-Systems der Google-Suche. Diese Mitwirkung ist relevant, um das KI-System dauerhaft zu betreiben.

Auch das ist ein Paradebeispiel für Human-Aided AI. Was wir hieran sehen ist, dass die Beteiligung von Menschen an solchen KI-Systemen *fortwährend* benötigt wird. Bei der Google-Bildersuche konnte man noch auf die (falsche) Idee kommen, dass der Bilder-Cache von Google irgendwann mittels des Browser-Spiels fertig

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

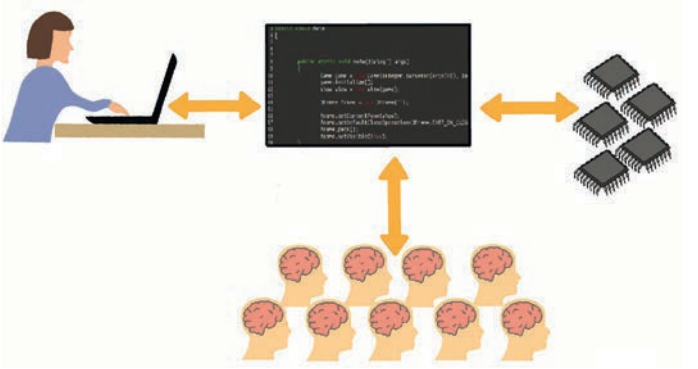


Abb. 5: Hybrides Prozessieren: Die Nutzer:in bedient am Frontend eine Anwendung, deren Berechnungen teils auf Siliziumprozessoren und durch Menschen ausgeführt werden. Grafik: Karla Baublys.

gelabelt ist, und dann benötigt man die menschliche Beteiligung nicht mehr.[13] Doch das ist nicht der Fall, denn ein ständig in Entwicklung befindlicher Kontext der realen Welt, auf die hier mittels KI zugegriffen werden soll, erfordert auch eine ständige Rekalibrierung des KI-Modells.

Das Beispiel der Google-Suche macht also klar: Menschliche Zuarbeit ist im Allgemeinen nicht nur ein initialer oder phasenweiser, sondern ein *integraler* Bestandteil von Human-Aided AI-Systemen. Die Rechenoperationen, die im Beispiel der Suchmaschine zur Pro-

13 Technisch steht hierbei die Idee im Hintergrund, dass das System vermeintlich von einer Phase des Trainings in eine »Inferenz«-Phase umschaltet. Diese Trennung von Training und Inferenz liegt aber im realen Einsatz von maschinellen Lernverfahren oft nicht vor. Auch im Fall der Google-Bildersuche ist die KI nicht irgendwann »fertig trainiert«, allein schon, weil ständig neue Bilder im Netz produziert werden, die wieder durch menschliche Mitarbeit gelabelt (oder hinsichtlich automatisch produzierter Labels überprüft) werden müssen.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

duktion von Suchresultaten auf eine Suchanfrage hin nötig sind, finden partiell auf Computerchips und partiell in menschlichen Gehirnen statt. Oder anders ausgedrückt: Das KI-System der Google-Suche ist ein Algorithmus, der sowohl Rechenoperationen auf siliziumbasierten Computerchips als auch das Auslesen menschlicher Reaktionen, Verhaltensweisen und Kognitionsleistungen orchestriert und zusammenführt (siehe Abbildung 5). Die Nutzer:innen eines solchen KI-Systems sitzen an einem Frontend-Interface – an einer grafischen oder haptischen Mensch-Maschine-Schnittstelle –, die von alledem wenig verrät. Im Hintergrund läuft ein Code, der die Suchanfragen verarbeitet, und dieser Code löst Rechenoperationen aus, die teilweise in Prozessorkernen in Rechenzentren ausgeführt werden, aber auch teilweise in menschlichen Gehirnen, indem irgendwo auf der Welt einer anderen Nutzer:in etwas angezeigt wird, worauf sie reagiert, zum Beispiel jener Nutzer:in, die zehn Minuten vorher eine ähnliche Google-Suche getätigt hat.

Das kritisch-philosophische Paradigma der Human-Aided AI etabliert also eine Gefügeperspektive auf KI-Systeme. Ein wenig kontinentalphilosophisch orientiert würde man sagen, diese KI-Systeme sind Agencements oder Assemblagen, also soziotechnische Gefüge. Die kritische Philosophie und Ethik der KI, die wir dringend benötigen, macht so eine Gefügeperspektive auf KI stark, um aufzuzeigen, wie sich soziale Praktiken und informatische oder elektrotechnische Vorgänge in Rechennetzen auf eine sehr komplexe Weise miteinander verschränken. Diese künstliche Intelligenzleistung kann ohne Bezug auf die soziale Komponente der Benutzung digitaler Services an vernetzten Endgeräten nicht adäquat

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

beschrieben werden. Solche Analysen von KI-Systemen sind komplex, denn das Spektrum von sozialen Praktiken bis zu den Schaltkreisen von Mikroprozessoren umfasst viele Ebenen und Bereiche: neue Formen von (impliziter, freier, prekärer) Arbeit und digitale Geschäftsmodelle, digitale Kultur und Subjektivität der Nutzer:innen in Bezug auf ihre Beziehung zu digitalen Geräten und Interfaces, politische, wissenschaftliche und gesellschaftliche Diskurse um KI und digitale Infrastruktur, Kapitalmarktdynamiken, in denen KI vielleicht überbewertet wird.

Die erste These von Human-Aided AI lautet deshalb, dass die meisten kommerziellen KIs heute hybride, gehirn- und siliziumbasierte Computernetze sind, die im Kontext gegenwärtiger Medienkultur ermöglicht werden. Das ist in etwa das, was Luis von Ahn mit seiner verheißungsvollen Formel gemeint haben dürfte: »We are going to consider all of humanity as an extremely advanced, large-scale distributed processing unit«.

D. Das Korrespondenzprinzip: Facebook-Foto-Tagging

Falls ich Sie noch nicht überzeugt habe, möchte ich den Blick auf ein weiteres lehrreiches Beispiel lenken. Das soziale Netzwerk Facebook bietet seinen Nutzer:innen seit dem Jahr 2010 die Funktion, auf hochgeladenen Fotos die Gesichter von Freund:innen zu markieren. Das heißt, der Bereich eines Fotos, auf dem ein Gesicht zu sehen ist, kann mit dem Facebook-Account der zugehörigen Person verknüpft werden (Abbildung 6). Tatsächlich gehört das Markieren von Gesichtern auf Fotos mittlerweile zum fest etablierten Repertoire sozialer Interaktionsformen auf vielen sozialen Medien. Es ist ja

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Abb. 6: Gesichter markieren auf Facebook. Grafik: Rainer Mühlhoff, Foto: Garry Knight, 2012, »Friends with Mobile Phones«. CC-BY 2.0. https://www.flickr.com/photos/garryknight/7003178857/in/photostream/.

auch eine schöne Funktion, sein Fotoalbum mit Hintergrundinformationen anzureichern oder die Namen all der Gesichter von der Party gestern Abend herausfinden zu können, die man sich nicht hat merken können.

Warum jedoch hat Facebook dieses Feature Ende der 2000er Jahre eingeführt? – Gewiss nicht allein, um die Welt mit besseren Fotoalben und Namensgedächtnisstützen auszustatten. Sondern Facebook hat damals für sich das Projekt formuliert, eine Gesichtserkennungs-KI bauen zu wollen – also ein KI-System, welches Gesichter auf Fotos erkennen kann. Was braucht man, um mittels maschineller Lernverfahren eine Gesichtserkennungs-KI zu bauen? Natürlich große Mengen von Trainingsdaten in Form von gelabelten Gesichtsbildern, also Gesichtsbildern, bei denen bekannt ist, wer darauf zu sehen ist. Denken Sie an das ESP-Game. Ein beliebtes soziales Netzwerk wie Facebook war in der besten Position, kos-

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

tenlos an einen solchen Datensatz zu kommen, wenn es seine Nutzer:innen dazu bringt, diese Daten freiwillig und als Teil der sozialen Interaktion auf der Plattform zu liefern.

Sie sehen hier eine Vorgehensweise, die dem ESP-Game sehr ähnlich ist, allerdings ohne das Element der »Gamification« auskommt. Das ESP-Game haben die Nutzer:innen relativ explizit als Spiel gespielt; es war hier auch kommuniziert, dass die dabei gewonnenen Daten zum Training von Bilderkennungs-KIs eingesetzt werden. Im Fall des Facebook-Foto-Taggings wird weniger ein Spieltrieb der Nutzer:innen ausgenutzt als vielmehr ihre sozialen Motivationen. Facebook hat eine Interaktionswelt erschaffen, in der es zum sozialen Miteinander gehört, Gesichter auf Fotos zu markieren. – Es ist übrigens überhaupt nicht selbstverständlich oder »natürlich«, dass Nutzer:innen sich hinsetzen und Gesichter auf Fotos markieren. Die sozial-mediale Interaktionswelt, in der das geschieht, ist vielmehr auf der Grundlage zielgerichteter Design-Überlegungen gestaltet; sie fällt, so wie sie ist, ja nicht vom Himmel.[14]

Facebook war im unentgeltlichen Einsammeln gelabelter Gesichtsbilder schließlich so erfolgreich, dass sie im Jahr 2017 das Projekt in das nächste Stadium überführen konnten. Sie hatten anhand des Datenmaterials eine KI trainiert, die ab diesem Jahr nicht mehr nur passiv lernte, sondern mittlerweile so treffsicher geworden war, dass sie von nun an dafür eingesetzt werden konnte, selbst aktiv Labels zu produzieren, wenn eine Person auf einem Foto nicht von einer Nutzer:in markiert wurde. Fortan galt also: Wenn Sie ein Bild hochluden, auf dem

14 Mühlhoff 2020a.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

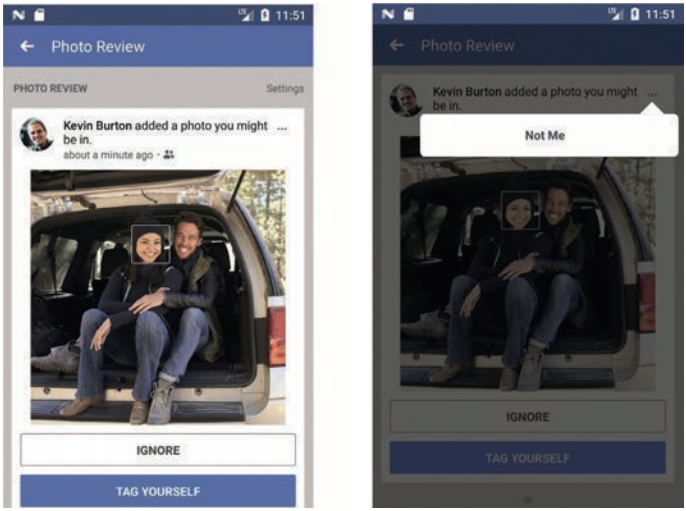


Abb. 7: Facebook benachrichtigt die Nutzer:in, dass ihr Gesicht auf einem hochgeladenen Foto erkannt wurde, und bietet drei Optionen zur Auswahl. Quelle: Facebook, 2017.

Sie eine bestimmte Person *nicht* markierten, dann konnte die Facebook-Gesichtserkennungs-KI seit 2017 solche Markierungen eigenständig vorschlagen.[15]

Dafür wurde eine ganz bestimmte Benutzerführung entworfen: Die Nutzer:in, die von der KI automatisch auf einem Foto gefunden wurde, bekam eine Nachricht (siehe Abbildung 7) darüber, dass jemand »ein Foto hochgeladen [hat], auf dem du zu sehen sein könntest«. Die Nutzer:in bekam dann die Möglichkeit, dazu Stellung zu nehmen, insofern dieser Dialogbildschirm sie vor verschiedene Auswahlmöglichkeiten stellte: Die Nut-

15 Nach anhaltender öffentlicher Kritik an den Privatsphäre- und Sicherheitsstandards auf der Plattform hat Facebook Anfang November 2021 angekündigt, den Einsatz von Gesichtserkennungstechnologie zur automatischen Markierung von Nutzer:innen au Fotos wieder einzustellen. Siehe Facebook 2021.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

zer:in konnte bestätigen, »Ja, das bin ich, und ich möchte auf dem Bild markiert werden«, oder »Ja, das bin ich, aber ich möchte auf dem Bild nicht markiert werden«. Wenn man in dem Dialog noch ein bisschen weitersuchte und auf das Menü-Symbol klickte, fand sich auch noch eine weitere Option: »Nein, das bin ich gar nicht«. Das heißt, falls es sich um eine fehlerhafte Zuordnung durch die KI handelte, konnte die Nutzer:in das ebenfalls zurückmelden.

Bemerkenswerterweise hat Facebook seinen Nutzer:innen die aktive Phase der Gesichtserkennungs-KI und die automatische Benachrichtigung, wenn man auf einem hochgeladenen Foto »zu sehen sein könnte«, als eine *Datenschutzfunktion* verkauft. In der Ankündigung für diese Funktion wurde suggeriert, sie sei dazu erschaffen worden, Nutzer:innen mehr Kontrolle über die Bilder zu geben, die von ihnen zirkulieren. Wir befanden uns hier, in den Jahren 2017 und 2018, längst in der Zeit nach den Snowden-Enthüllungen und – noch aktueller – in der Zeit direkt nach dem Cambridge-Analytica-Skandal, der Facebook in Bezug auf Datenschutz stark unter Druck gesetzt hat. Eine Sensibilität für Privatsphäre-Themen, gerade in Bezug auf das soziale Netzwerk, war da weit verbreitet. An der Werbestrategie für das Gesichtserkennungsfeature sieht man, wie große Unternehmen diese Privacy-Sensibilisierung für sich auszunutzen und umzudeuten versuchten.

Doch hat Facebook die Gesichtserkennungs-KI, die ein langjähriges Projekt gewesen war, dessen Anfänge weit vor die Zeit der Snowden-Enthüllungen zurückdatieren, wirklich als Maßnahme zur Verbesserung der Kontrolle über die eigenen Daten entwickelt und eingeführt? Es spricht einiges dafür, dass die interne Logik

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

hier eine ganz andere war. Es liegt nahe, dass die oben beschriebenen Benachrichtigungsdialoge, in denen Nutzer:innen entscheiden sollten, ob sie auf einem Bild markiert werden möchten oder nicht, und ob es sich überhaupt um sie selbst handelt, in der Logik des KI-Systems einem ganz anderen Zweck dienten: Dieses KI-System war 2017 in eine aktive Phase eingetreten, in der es begann, selbst Markierungen auf Fotos zu produzieren. Und dabei war es auf *Verifikationsdaten* angewiesen – auf einen Dateninput von Menschen, die die Markierungen der KI überprüfen und gegebenenfalls korrigieren. Die ersten beiden Auswahlmöglichkeiten – ob die Nutzer:in markiert werden oder inkognito bleiben möchte – machen für das KI-System keinen Unterschied, denn sie beide bestätigen, dass es sich um die Person selbst handelt, die automatische Erkennung also erfolgreich war. Relevant sind für das KI-System die ersten beiden Optionen in Abgrenzung zur dritten – der Möglichkeit einer falschen Erkennung. Beide Daten, die richtige und die falsche Erkennung, können, sobald sie von Menschen bestätigt wurden, als Trainingsdaten wieder in das System eingespeist werden, um das KI-Modell ständig zu verbessern und einer sich verändernden Welt anzupassen.

Sie sehen, dass ein Teil der »Intelligenz« dieses KI-Systems durch eine spezifische Gestaltung des Bestätigungsdialogs zustande kommt – also durch Know-how auf der Ebene des *Designs* von Benutzerschnittstellen. Das Design dieses Interfaces zielt darauf ab, Nutzer:innen ohne ihr Wissen dafür einzuspannen, Trainings- und Verifikationsdaten für diese KI zu produzieren. Ein wesentlicher, funktionaler Teil dieses KI-Systems ist durch Expertise auf dem Gebiet des Designs von Mensch-

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Maschine-Interaktion unter Ausnutzung eines bestimmten gesellschaftlichen Diskurses (Sensibilisierung für Privatsphäre) umgesetzt worden. Wir sehen: KI-Systeme werden nicht nur programmiert, sondern auch die Künste von Designer:innen fließen in solche Projekte ein.

Das führt mich schließlich zur zweiten These des Human-Aided AI-Paradigmas, die ich das Korrespondenzprinzip nenne. Wenn Sie ein bisschen die Industrie studieren, dann sehen Sie die Tendenz, dass sich Probleme des *machine learning* in Probleme des Designs von Mensch-Maschine-Interaktion übersetzen lassen. Probleme im Bereich KI (Sie möchten ein bestimmtes Machine-learning-Modell trainieren) und Probleme im Bereich Interfacedesign (Sie möchten die Nutzer:innen dazu bringen, Ihnen kostenlos die Trainingsdaten zu liefern) stehen in einer Korrespondenzrelation.

E. Eine Mediengeschichte der Trainingsdatenproduktion

Wenn man diese Korrespondenz zwischen Interfacedesign und KI ausführlich studiert, kann man eine Art Timeline erstellen, in der die Korrespondenz von Entwicklungen im Bereich Mensch-Maschine-Interaktion und KI in ihrer historischen Tiefe sichtbar werden (vgl. Tabelle 1). In den 1990ern wurde viel vom sogenannten *ubiquitous computing* (kurz: *ubicomp*) gesprochen. Dies war eine Prophetie des Informatikers und Silicon-Valley-Visionärs Mark Weiser, der so etwas wie die Allgegenwart von Computern, elektronischen Sensoren und rechnergestützter Informationsverarbeitung kommen sah.[16] Digitale Datenverarbeitung und Sensorik würden in jede Nische

16 Weiser 1991.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

des Lebens Einzug halten, in jede Hosentasche, jeden Kühlschrank, jeden privaten, wirtschaftlichen oder öffentlichen Vorgang. Spätestens seit den 2000er Jahren sehen wir deutlich, dass er recht hatte. Digitale Medieninfrastruktur, vernetzte Sensoren und Rechengeräte, das Internet der Dinge etc. verbreiten sich in globalem Maßstab und dringen zugleich in jeden kleinsten Lebensbereich ein.

~1991	Ubiquitous computing (Mark Weiser)
⋮	
1997	Google-Suche
2003	ESP Game
~2004	Web 2.0
2005	Amazon Mechanical Turk Google Analytics
2006	Facebook
~2010 ff.	Smartphone
⋮	
~2016 ff.	Deep-Learning-Hype

Tab. 1: Wichtige technische und konzeptuelle Meilensteine in der Entwicklung vernetzter digitaler Medien. Im ersten Jahrzehnt des 21. Jahrhunderts entsteht die Medieninfrastruktur zur Produktion von KI-Trainingsdaten.

Hierbei sehen wir aber gerade in den 2000er Jahren eine Schlüsselentwicklung, die Mark Weiser zwar bestätigt, die aber zugleich ein wichtiges Merkmal zeigt, das über seine Vorhersage hinaus geht: Was sich stark verbreitet, das ist vor allem *vernetzte* digitale Medieninfrastruktur. Es handelt sich nicht einfach um Rechengeräte in unseren Hosentaschen und um »intelligente« Kühlschränke, sondern vor allem um *vernetzte* Kommunikationsmedien in den Hosen-

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

taschen und *vernetzte* Küchengeräte. Es zeigt sich, dass jene digitalen Geräte und Dienste, die seit den 2000er Jahren die meiste Verbreitung gefunden haben und wirtschaftlich erfolgreich (das heißt auch: finanziell tragfähig) waren, genau diejenigen sind, die die Extraktion von Trainingsdaten für KI-Systeme ermöglichen. Der intelligente Kühlschrank selbst lässt sich nicht so gut monetarisieren wie der Kühlschrank, der seine Daten in ein Netzwerk zur weiteren Verwendung einspeist; das Smartphone und seine Apps werden auch erst durch die Anbindung an die Cloud zu einem expansiven Geschäftsmodell. Datenaggregation ist somit die Bedingung der ökonomischen Realität von *ubicomp* im 21. Jahrhundert. Diese Realität zeigt sich an Meilensteinen wie der Entwicklung der Google-Suche Ende der 1990er Jahre; dem Web 2.0-Paradigma ab ca. 2004; der Entstehung der Klickarbeit in verschiedenen Formen wie zum Beispiel sichtbar im ESP-Game oder bei Amazon Mechanical Turk in denselben Jahren; der Etablierung weiterer Tracking-Infrastrukturen wie mittels Cookies oder Google Analytics; der Konjunktur der sozialen Medien und vernetzten Telefone seit Mitte/Ende der 2000er Jahre (siehe Tabelle 1).

So waren wir in den 2000er Jahren Zeug:innen einer technischen Entwicklung, die in ihrer Summe eine ubiquitäre und fein differenzierte vernetzte Infrastruktur für die Aggregation oder das *mining* von weitestgehend kostenlosen Datenschätzen hervorgebracht hat. Der Trend der Computarisierung jeder Lebensnische ist gekoppelt an einen großen Trend der Aggregation und zentralen Zusammenführung all der dabei entstehenden Daten. Diese Datenschätze, die bei der Industrie liegen, machten mit einer Latenz von etwa 10 Jahren, also grob seit den 2010er Jahren, Deep Learning und andere stark

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

datenbasierte KI-Verfahren allererst möglich. Es ist kein Zufall, dass wir in den 2000ern die Verbreitung dieser Technologien sehen und in den 2010ern plötzlich den Hype um Deep Learning – denn Deep Learning war nur möglich, *weil* es diese Daten gab.[17]

Es ist entscheidend, solche medienhistorischen – genauer müsste man sagen: medien-genealogischen – Untersuchungen nicht mit Verschwörungsnarrativen zu verwechseln. Die Behauptung ist hier nicht, dass einzelne Akteure und Profiteure diese Entwicklung von langer Hand geplant und konspirativ herbeigeführt hätten. Es handelt sich um eine Entwicklung ohne Zentrum und ohne Mastermind, die dennoch im Zusammenspiel vieler Akteure und Interessen eine strategische Qualität[18] besitzt, gesellschaftliche Transformationen und bestimmte Profiteure hervorbringt. Zu den Produkten dieser Transformation gehören die neuen vernetzten digitalen Medien und eine neue Welle von Erfolgen auf dem Gebiet der künstlichen Intelligenz – diese Welle bevorzugt nun die datenbasierten Ansätze in der KI und nicht die logisch-deduktiv orientierten KI-Paradigmen, die in der zweiten Hälfte des 20. Jahrhunderts stark betont wurden.[19]

An dieser kleinen Geschichte der Trainingsdatenproduktion zeigt sich, wie das kritisch-ethische Paradigma der Human-Aided AI auch eine historische Dimension gewinnt und die Frage nach der Entstehung des Status quo der KI-Technologie mit Blick auf ihre soziotechnischen Voraussetzungen erhellen kann. Insbesondere

17 Siehe ausführlicher Mühlhoff 2019b.

18 »Strategisch« verstehe ich hier im Sinne von Michel Foucault, vgl. Foucault 1983 [1976]: 93–95.

19 Siehe zum Beispiel Haugeland 1985.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

zeigt sich, dass die Geschichte der KI als eine Mediengeschichte erzählt werden kann und – im Sinne eines kritischen Projekts – auch als solche erzählt werden *sollte*. Denn genauso wie die sichtbaren Erfolge von KI, die wir alle verwenden und auf die wir täglich zugreifen können, gehört auch ein neuer Grad der Vernetzung nahezu aller Menschen auf diesem Planeten und der dadurch ermöglichten Datenaggregation zu diesem aktuellen »Sommer der KI-Erfolge«.[20] Es kommt darauf an, die sozioökonomischen Effekte dieser Entwicklung nicht als Nebeneffekt, sondern als integralen Teil von dem zu begreifen, was heute mit »künstlicher Intelligenz« gemeint ist.

Teil 2: KI zur Vorhersage persönlicher Informationen

In der beschriebenen soziotechnischen Perspektive auf KI möchte ich den Fokus nun ein Stück verengen und auf eine bestimmte, zurzeit weit verbreitete Anwendung von KI-Technologie schauen: nämlich auf KI-Systeme zur Vorhersage unbekannter und oft persönlicher Informationen. Ich verwende dafür den Sammelbegriff der *prädiktiven Analytik*. Diese Technologie ist exemplarisch für jene Anwendungsdomänen von KI, von denen die meisten Menschen nichts wissen. Überdies geht diese Technologie mit ethischen Problemen einher, für die unsere wissenschaftlichen Debatten noch nicht gut genug gerüstet sind.

Um direkt ein Beispiel für prädiktive Analytik zu benennen, können wir auf ein allbekanntes Paper von Michal Kosinski et al. aus Cambridge schauen.[21] Dort

20 Ng 2017.

21 Kosinski u. a. 2013. Siehe darin insbesondere Figure 2.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

wird gezeigt, dass man aus einer Handvoll Facebook-Likes eine Reihe von persönlichen Informationen über Facebook-Nutzer:innen ableiten kann. Sind Sie Single oder in einer Beziehung? Sind Sie mit getrennten Eltern aufgewachsen? Rauchen Sie, missbrauchen Sie Alkohol oder andere Substanzen? Welche ethnische oder religiöse Zugehörigkeit haben Sie? Welche politischen Einstellungen haben Sie? Sind Sie homosexuell? Was ist Ihr Geschlecht? Ein anderes Paper hat vor wenigen Jahren gezeigt, dass man aus dem Inhalt von Social-Media-Postings detaillierte Vorhersagen über Krankheiten wie Diabetes, Bluthochdruck oder STIs ableiten kann.[22]

All diese sensiblen Informationen lassen sich aus den Nutzungsdaten der betroffenen Person auf Social-Media-Plattformen vorhersagen, also aus den Daten darüber, welche Inhalte wir anschauen, liken, teilen oder bookmarken. Es ist dafür nicht erforderlich, dass die betroffene Person die vorhergesagten Informationen irgendwo anders angegeben hat. Facebook als Plattform ist hierfür natürlich nur ein Beispiel, das Gleiche gilt für alle anderen sozialen Medienplattformen. Prädiktive Analytik funktioniert außerdem auch mit anderen Nutzungsdaten, die außerhalb von Plattformen gesammelt werden können, wie etwa dem Browserverlauf oder den Daten von Tracking-Cookies. Trackingdaten im Internet werden zum Beispiel dafür verwendet, Versicherungsrisiken, Kreditwürdigkeit, die persönliche Kaufkraft, den Bildungsgrad, Suchtdispositionen, persönliche Interessen, demografische Informationen, Geschlecht und unseren Lebensstandard einzuschätzen.

22 Merchant u. a. 2019.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Prädiktive Analytik ist als Begriff nicht spezifisch in Bezug auf die verwendeten Algorithmen. Das können maschinelle Lernverfahren verschiedener Art sein, aber auch einfachere Regressionsanalysen oder statistische Auswertungen. Für die folgenden Ausführungen ist das im Detail nicht entscheidend, ich gehe lediglich davon aus, dass das Verfahren aus empirischem Datenmaterial lernt, seine Vorhersagen zu erstellen.

Relevant für die ethische Besprechung ist somit eine *funktionale* Charakterisierung prädiktiver KI-Systeme (siehe Abbildung 8):[23] Ein prädiktives Modell ist eine »Maschine«, die als Input die bekannten Daten über ein Individuum oder einen Fall erhält (z. B. Trackingdaten oder Facebook-Likes) und als Output eine Schätzung unbekannter Informationen über das Individuum oder den Fall ausgibt. Die große Frage ist natürlich: Wie macht ein prädiktives Modell das? Man kann aber auch anders fragen: *Wer* kann eigentlich so etwas machen? *Wer* kann solche Maschinen herstellen, die anhand von leicht verfügbaren Daten schwer zugängliche oder höchst sensible Informationen abschätzen? – Das können große Datenunternehmen, zum Beispiel eine große Social-Media-Plattform, die sehr viele Nutzer:innen hat, die täglich Daten hinterlassen – nehmen wir doch als Beispiel Facebook. In Abbildung 8 entsprechen die grünen Kästchen dann gesammelten Facebook-Likes jeweils einer Plattform-Nutzer:in. Ein paar der Nutzer:innen hinterlassen nicht nur Likes auf der Plattform, sondern geben auch bestimmte persönliche oder sensiblen

23 Ich folge hier der Darstellung in Mühlhoff 2022, 2021.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

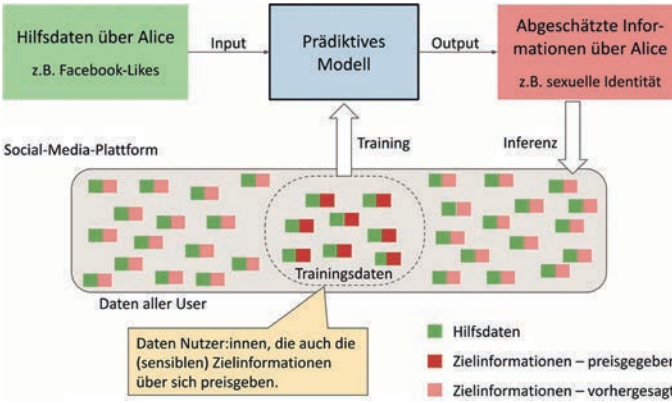


Abb. 8: Funktionsweise prädiktiver Analytik im Kontext von Social-Media-Daten. Grafik: Rainer Mühlhoff.

Informationen über sich preis; beispielsweise machen einige Nutzer:innen in ihrem Facebook-Profil Angaben zu ihrer sexuellen Orientierung. Solche Informationen sind in Abbildung 8 als rote Kästchen eingetragen. Es handelt sich bei den roten Datenpunkten um die Zielvariable eines prädiktiven Modells, das anhand dieser Daten erstellt werden kann – in unserem Beispiel wäre das ein prädiktives Modell zur Abschätzung der sexuellen Orientierung anhand von Facebook-Likes.

Gehen wir einmal davon aus, dass nur wenige Prozent der Nutzer:innen einer Social-Media-Plattform dazu bereit sind, explizite Angaben über ihre sexuelle Orientierung zu machen. Dann sind das im Fall von Facebook dennoch einige Hundert Millionen Nutzer:innen. Von dieser Untergruppe der Nutzer:innen – in Abbildung 8 sind sie zu dem umkreisten Bereich in der Mitte zusammengefasst – sind also die Hilfsdaten *und* die Zielinformationen verfügbar. Dieser kombinierte Datensatz kann als sogenannter »Trainingsdatensatz« für ein

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

maschinelles Lernverfahren verwendet werden, welches in diesen Daten automatisch nach Korrelation zwischen den Hilfsdaten und den sensiblen Daten sucht. Wenn Sie also ein großes Social-Media-Unternehmen sind, dann bauen Sie prädiktive Modelle aus den Daten *derjenigen Minderheit Ihrer Nutzer:innen*, die Ihnen die sensiblen Zielinformationen freiwillig überlassen.

Sobald das Plattformunternehmen aus den Daten dieser »freigiebigen« Minderheit ein prädiktives Modell für sexuelle Orientierung trainiert hat, ist es dazu in der Lage, dieses Modell auf *alle anderen* Nutzer:innen anzuwenden – also auf die, die in Abbildung 8 nur mit einem grünen Kästchen verzeichnet sind. Die schwach-roten Kästchen repräsentieren in der Abbildung diese im Nachgang vorhergesagten Daten über die sexuelle Identität dieser Nutzer:innen. Das anhand der Daten der »freigiebigen« Minderheit trainierte Modell erlaubt es der Plattform also, die sensiblen Informationen auch über jene Nutzer:innen abzuschätzen, die diese Daten nicht explizit angegeben haben und potenziell auch nicht angeben wollen. Wenn Sie also eine Social-Media-Plattform anonym nutzen oder bestimmte Informationen gezielt *nicht* angeben, können Sie davon ausgehen, dass die Plattform diese Informationen trotzdem über Sie ermittelt. Plattformen interessieren sich sehr für solche prädiktiven Modelle, denn die damit abschätzbaren Informationen über beliebige ihrer Nutzer:innen lassen sich lukrativ weitervermarkten: in der Finanz- und Versicherungsbranche, in der Werbeindustrie, im Bereich der Personalführung etc.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Shoshana Zuboff, eine der großen kritischen Theoretiker:innen der prädiktiven Verwendung von Big Data und KI, spricht vom sogenannten »Überwachungskapitalismus«. Damit meint sie eine in den letzten 10 bis 20 Jahren entstandene Form der Kapitalakkumulation, die auf der Ausbeutung von Nutzungsdaten im Internet beruht. Die Geschäftsmodelle vernetzter digitaler Dienste sind so konstruiert, dass möglichst viele Daten über uns erfasst werden – über unser Verhalten, unsere Präferenzen, unseren Aufenthaltsort, unsere Kommunikationsbeziehungen, unsere E-Mails, Textnachrichten, Suchanfragen etc. Das geht so weit, dass zum Beispiel Automobilhersteller ihre Fahrzeuge nicht mehr allein und primär als Fortbewegungsmittel sehen, sondern als fahrende Sonden zur Vermessung von Welt und Menschen.[24]

Die Anbieter kostenloser digitaler Dienste arbeiten in der Regel mit mindestens zwei Geschäftsmodellen: Das eine ist ein Frontend- und das andere ein Backend-Geschäftsmodell. Am Frontend wird ein Produkt angeboten, das bestechend praktisch ist und das Sie gerne benutzen möchten, sagen wir die Google-Suche, Google Docs, Google Maps, YouTube, Instagram, Apples Spracherkennung etc. An diesem Frontend sind Sie Nutzer:in, nicht Kund:in. Erst am Backend werden die zahlenden Kund:innen dieser Unternehmen gewonnen – und das sind nicht die Nutzer:innen. Diese Kund:innen kaufen vielmehr dort Zugriff auf *Sie*. Zum

24 Zuboff 2015.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

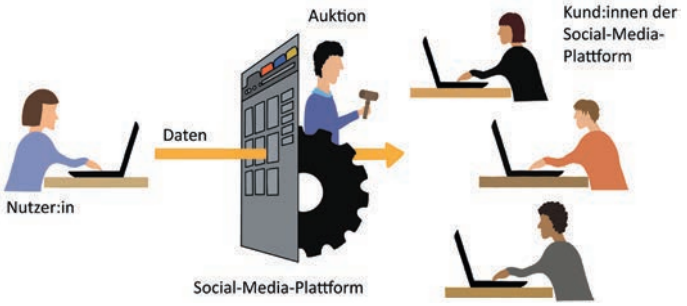


Abb. 9: Frontend- vs. Backend-Geschäftsmodell. Kund:innen kaufen nach einem Auktionsverfahren Zugriff auf Vorhersageprodukte, die aus Nutzungsdaten erstellt werden. Grafik: Karla Baublys.

Beispiel um zu erfahren, wie riskant Sie als Versicherungsnehmer:in sind oder auf was für eine Werbeanzeige Sie heute, in Ihrer aktuellen Stimmungslage und vor dem Hintergrund Ihres Persönlichkeitsprofils mit der größten Wahrscheinlichkeit klicken werden. Die im Frontend-Geschäft gewonnenen Nutzungsdaten dienen den Unternehmen also dazu, »Vorhersageprodukte« herzustellen und in anderen Marktsegmenten zu verkaufen. Damit kann man mehr Gewinn erzielen als mit einer direkten Monetarisierung der Frontend-Produkte zum Beispiel über eine Nutzungsgebühr. Die Frontend-Produkte werden also nur indirekt über den Verkauf prädiktiver Modelle in anderen Marktsegmenten monetarisiert. Es handelt sich dabei um jene Marktsegmente, von denen ich bereits in der Einleitung sprach: Credit Scoring und das Scoring von Versicherungsrisiken, Automatisierung von Auswahlverfahren für Jobs oder Studienplätze, Personalführung, individuelle Preisgestaltung für Produkte aller Art, individualisierte Werbung, Produktvorschläge, algo-

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

rithmische Unterstützung des Jugendschutzes, der Polizei, der Justiz usw.[25]

In vielen Backend-Märkten wurden für den Zugriff auf Vorhersageprodukte von den großen Plattformunternehmen Auktionsmechanismen implementiert. Das heißt, die Meistbietende kann die Vorhersagen kaufen (Abbildung 9). Es geht dabei um Gewinnmaximierung; moralische oder politische Erwägungen spielen fast keine Rolle. Als Nutzer:in sehen Sie nie etwas von diesen Backend-Märkten, weil diese Ebene der Geschäftsaktivität der Plattformunternehmen eben hinter der Oberfläche der bunten Nutzerinterfaces stattfindet. Die Backend-Geschäftsaktivitäten dieser Unternehmen muss man sich hinsichtlich ihres Umfangs allerdings eher wie den Eisberg unter der Wasseroberfläche vorstellen: Verglichen mit dem, was ein schlichtes, minimalistisches, öffentlich sichtbares Frontend-Interface suggeriert, sind sie deutlich umfangreicher.

C. Psychologisches Targeting und Wahlwerbung

Spätestens seit dem Facebook-Cambridge-Analytica-Skandal im Nachgang der Wahl Donald Trumps im Jahr 2016 kennen wir aus der medialen Debatte das Prinzip des prädiktiven Targetings von Wahlwerbung. Das mittlerweile insolvente Datenanalyse- und Politikberatungsunternehmen Cambridge Analytica hatte aus Facebook-Daten prädiktive Modelle zur Abschätzung von psycho-

25 Hier einige detailliertere Studien zu den genannten Anwendungsgebieten prädiktiver Analytik: O'Neil 2016; Eubanks 2017; Keddell 2015; Lippert 2014; Wagner und Eidenmuller 2019; Martini u. a. 2020.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

logischen Charaktereigenschaften erstellt, die dann zur Beeinflussung von Wahlen durch sogenanntes Mikrotargeting verwendet wurden. Der ehemalige CEO von Cambridge Analytica, Alexander Nix, behauptet, seine Firma sei allein in den USA in mehr als 40 Wahlkämpfe involviert gewesen, unter anderem in die Kampagnen von Ted Cruz und später Donald Trump bei den US-Präsidentschaftswahlen 2016, außerdem wird Cambridge Analytica in Zusammenhang gebracht mit der Leave.eu-Kampagne beim britischen Brexit-Referendum im selben Jahr.[26] Eine der Datengrundlagen der Modelle von Cambridge Analytica bildeten Daten über ca. 80 Millionen Facebook-Nutzer:innen, die von Aleksandr Kogan, damals ein Psychologe an der Universität Cambridge, im Rahmen »akademischer« Untersuchungen über Facebook-Nutzer:innen erhoben wurden.[27] Kogan agierte dabei für das von ihm gegründete Unternehmen Global Science Research (GSR), welches diese Daten mit einer Facebook-App, also einer in Facebook eingebetteten Software des Unternehmens, die eine Art psychologisches Quiz darstellte, sammelte und später an die SCL Group, den Mutterkonzern von Cambridge Analytica, verkaufte.

Welche Modelle Cambridge Analytica wirklich anhand solcher Daten trainieren konnte, ist nicht abschließend bewiesen. Eine plausible und insbesondere von Alexander Nix selbst verbreitete Version der Geschichte besagt, dass die Vorhersagemodelle eine Abschätzung psychologischer Charaktereigenschaften nach

26 Sellers 2015; Hern 2018; Confessore und Hakim 2017.
27 Rosenberg u. a. 2018; Grassegger und Krogerus 2016.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Abb. 10: Zwei Werbebotschaften der Ted-Cruz-Kampagne, die Cambridge Analytica CEO Alexander Nix in einem Vortrag im Jahr 2016 vorstellte. Quelle: Cambridge Analytica, https://www.youtube.com/watch?v=n8Dd5aVXLCc.

dem OCEAN-Modell der empirischen Persönlichkeitsforschung vornahmen (im Deutschen häufig Fünf-Faktoren-Modell genannt). Dabei werden Menschen in den fünf Dimensionen »Aufgeschlossenheit« (*openness*), »Gewissenhaftigkeit« (*conscientiousness*), »Extraversion« (*extraversion*), »Verträglichkeit/Rücksichtnahme/Kooperationsbereitschaft« (*agreeableness*) und »Neurotizismus / emotionale Labilität / Verletzlichkeit« (*neuroticism*) eingeordnet.[28] Die Idee ist nun, dass Cambridge Analytica Schätzwerte dieser Charaktermetriken mittels geeigneter prädiktiver Modelle aus Facebook-Nutzungsdaten beliebiger Nutzer:innen gewinnen konnte. Solche Vorhersagemodelle konnten mutmaßlich mittels der aus der Kogan-Forschung erworbenen psychologischen Daten von 80 Millionen Facebook-Nutzer:innen in Kombination mit Plattform-Nutzungsdaten der Individuen trainiert werden (grüne und rote Datenpunkte in Abbildung 8). Cambridge Analytica war also mutmaßlich dazu in der

28 Matthews u. a. 2003, 23–24.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Lage, über alle Facebook-Nutzer:innen ein detailliertes psychometrisches Profil abzuschätzen.

Hat man einmal ein solches Vorhersagemodell für psychologische Charaktereigenschaften erstellt, dann folgt die Frage, wie sich damit Wahlen beeinflussen lassen. In Bezug auf die Kampagne von Ted Cruz in den Vorwahlen der US-amerikanischen Präsidentschaftswahl 2016 wurde bekannt, dass Cambridge Analytica auf verschiedene psychologische Persönlichkeitstypen jeweils individuell angepasste Werbebotschaften entworfen hat (siehe Abbildung 10).[29] Ein psychologisches Vorhersagemodell, welches anhand von Facebook-Nutzungsdaten die Eingruppierung in die relevante Persönlichkeitsgruppe vornimmt, konnte dann dazu verwendet werden, jeder Nutzer:in automatisiert die für ihr Persönlichkeitsprofil »passende« Version des Werbematerials anzuzeigen. So erläuterte der Cambridge-Analytica-CEO Alexander Nix in einem Vortrag in Bezug auf die Ted-Cruz-Kampagne, dass man Wähler:innen für eine Erhaltung des zweiten Verfassungszusatzes (Recht auf Waffenbesitz) mobilisieren wollte. Dazu sei den als »neurotic« und »conscientious« klassifizierten Individuen die in Abbildung 10 links stehende Werbebotschaft, den als »closed« und »agreeable« klassifizierten Individuen die rechts abgebildete Version angezeigt worden.

Man nennt eine solche Vorgehensweise psychologisches Targeting, denn die Zielgruppe einer Werbemaßnahme wird nach psychologischen Merkmalen eingeteilt und mit zugeschnittenen Botschaften adressiert. Dieses Verfahren fällt in die allgemeinere Kategorie des

29 Confessore und Hakim 2017; Davies 2015.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Mikrotargetings, einer Kommunikationsstrategie, welche im Kontext von Wahlen möglichst präzise umrissene (Kleinst-)Gruppen identifiziert, bei denen es sich lohnt, mit maßgeschneiderten Botschaften ihr Verhalten zu beeinflussen (zum Beispiel unentschlossene Wähler:innen in Swing States). Durch das Definieren möglichst kleiner Gruppen können Ressourcen für Werbung sehr effizient eingesetzt werden, da nur diese Personen und nicht eine breite Öffentlichkeit mit Botschaften versorgt wird. Außerdem können dadurch verschiedene Targeting-Gruppen in der oben beschriebenen Weise mit verschiedenen Botschaften versorgt werden, um öffentliche Gegenreaktionen auf besonders polarisierende Werbebotschaften zu vermeiden.

Dass Mikrotargeting im Kontext US-amerikanischer Wahlen eingesetzt wird, ist kein neues Phänomen. Bereits die Barack-Obama-Kampagnen in den Jahren 2008 und 2012 haben so agiert – man kann sogar sagen, dass sie Vorreiter dieser Techniken war, insofern Mikrotargeting durch sie erst weltweit bekannt wurde.[30]

Was macht den Facebook-Cambridge-Analytica-Skandal dann so dramatisch? Es zeigt sich daran eine neue Qualität. Diese Qualität liegt darin, dass mutmaßlich die Daten einiger Zehn Millionen Facebook-Nutzer:innen dafür verwendet wurden, prädiktive Modelle für psychologische Merkmale zu trainieren, die flächendeckend auf die gesamte Wahlbevölkerung angewendet werden konnten, um sie nach psychologischen Merkmalen zu gruppieren und einige von ihnen mit zugeschnittener Werbung zu versorgen. Entscheidend ist hier, wie ich im nächsten Abschnitt noch ausführlicher

30 Tufekci 2014.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

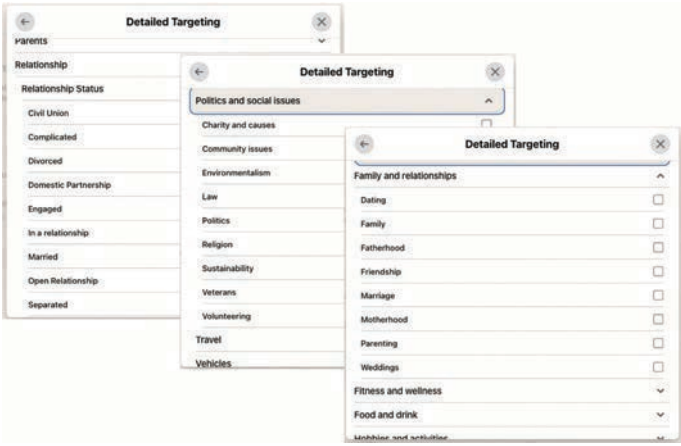


Abb. 11: Targeting-Kriterien für Facebook-Werbung (Auszug). Screenshots des Autors, facebook.com, 2021.

behandeln werde, dass die Daten, die von einigen (potenziell anonymisiert) erhoben wurden – nämlich von denen, die *freiwillig* in einem Psychologie-Quiz spielerisch ihren Charakter haben vermessen lassen –, dazu verwendet werden können, solche Informationen über *viele andere* Individuen abzuschätzen, die nichts davon wissen und die nicht aktiv bei einem Quiz mitgemacht haben.

Das Prinzip des prädiktiven psychologischen Targetings, das wir aus dem Facebook-Cambridge-Analytica-Skandal kennen, hat viele Debatten zur Beeinflussung der Demokratie durch soziale Medien und ihre KI-basierten Datenanalysemöglichkeiten ausgelöst. Wichtig ist, dass das Problem tatsächlich viel größer ist als der Facebook-Cambridge-Analytica-Skandal – schließlich ist Cambridge Analytica längst insolvent. Das Phänomen ist nicht auf Cambridge Analytica beschränkt, denn große Unternehmen wie Facebook, Google, Amazon und Apple besitzen alle relevanten Daten, um selbst prädiktive Modelle für das Mikrotargeting von (Wahl-)Werbung

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

zu trainieren. Um kurz bei Facebook zu bleiben: Abbildung 11 zeigt Screenshots aus dem Facebook-Backend für Werbekund:innen. Das ist ein kleiner Ausschnitt der von Facebook allgemein angebotenen Targeting-Kriterien, anhand derer Sie Ihre Zielgruppe definieren können, wenn Sie Werbung auf Facebook schalten möchten. Sie können die Nutzer:innen der Plattform zum Beispiel nach Geschlecht, Alter, Beziehungsstatus, politischer Orientierung und vielen weiteren detaillierten Kriterien adressieren. Hier sind auch zahlreiche Datenfelder dabei, die die meisten Menschen über sich nicht angeben würden. In solchen Fällen schätzt Facebook diese Informationen anhand prädiktiver Modelle über die Nutzer:in ab. Sie brauchen also für manipulative Datenverwendung dieser Art gar keine externe Firma wie Cambridge Analytica; und das eigentlich Skandalöse am Facebook-Cambridge-Analytica-Skandal ist auch *nicht*, dass eine externe Firma diese Daten erhalten hat. Sondern das Kernproblem ist, dass wer auch immer aggregierte Social-Media-Daten besitzt – und zunächst sind das die Plattformunternehmen selbst –, detaillierte prädiktive Modelle mit diesen Daten trainieren kann. Google, Apple, Facebook und Amazon sind alle im Geschäftsfeld des Mikrotargetings und der individualisierten Werbung aktiv.

D. Ist Targeted Advertising nicht in den meisten Fällen harmlos?

Wir haben jetzt erläutert, dass maßgeschneiderte Werbung manipulativ sein kann, insbesondere wenn sie Vorhersagen über psychologische oder emotionale Dispositionen der Nutzer:innen ausnutzt. Diese Form der gezielten Beeinflussung kann im Fall von Wahlwerbung für den

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Wahlausgang entscheidend sein, insbesondere wenn damit gezielt Personengruppen in Kontexten adressiert werden, wo wenige Stimmen einen großen Unterschied machen (Tipping-point-Konstellationen, Swing States etc.).

Doch wie würde unsere ethische Bewertung des Mikrotargetings ausfallen, wenn wir die *manipulative* Verwendung von prädiktiven Modellen ausschließen könnten? Nehmen wir einmal an, Mikrotargeting für Wahlwerbung wäre verboten – genauso wie das Targeting nach psychologischen und emotionalen Kriterien. Wäre individualisierte Werbung – etwa für Produkte oder Veranstaltungen – dann noch immer ethisch bedenklich? Eine aktuelle Werbekampagne von Facebook, die Nutzer:innen von dem vermeintlichen Mehrwert personalisierter Werbung überzeugen soll, bringt das alte Argument neu vor, dass die Nutzer:innen durch personalisierte Werbung mehr für sie relevante und interessante Werbung sehen und weniger von generischen Werbeanzeigen gestört würden.[31] Würde dieser Vorteil gegenüber den ethischen Bedenken nicht tatsächlich überwiegen? Hätten wir es bei Targeted Advertising – innerhalb der genannten, hypothetisch gesetzten Grenzen – nicht mit einer der harmlosesten Anwendungen prädiktiver Modelle zu tun?

Die Antwort lautet: *Nein, ganz und gar nicht.* Und ich möchte ein weiteres Beispiel verwenden, um Ihnen das zu erläutern.[32]

Eine Anwendungsdomäne, in der man sich viel von individualisierter Werbung verspricht, ist die Forschung über neue Medikamente und Therapieansätze in soge-

31 Vgl. exemplarisch Hutchinson 2022.

32 Die folgende Darstellung beruht auf der Studie Mühlhoff und Willem 2022.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

nannten klinischen Studien. Das sind groß angelegten Studien mit Proband:innen, in denen neue Medikamente oder Therapien ausprobiert werden. Stellen Sie sich vor, Sie möchten eine neue Therapie zur Behandlung von Diabetes Typ 2 (kurz: D2M) testen. Anstatt beispielsweise in der U-Bahn, in Arztpraxen oder illustrierten Magazinen Werbung für Ihre Studie zu platzieren, schalten Sie Werbung auf Social-Medial-Plattformen. Über die Targeting-Kriterien grenzen Sie die Zielgruppe direkt sinnvoll ein. Proponenten dieser Methode argumentieren, dass man so für das gleiche Werbebudget mehr relevante Adressat:innen erreichen kann als bei breit gestreuter Werbung, und im Fall von medizinischer Forschung würde diese Kosteneffizienz letztlich uns allen zugute kommen.[33] Abbildung 12 zeigt eine solche Werbung für eine D2M-Studie, die auf Facebook geschaltet wurde. Da können Sie sich sogar direkt über Facebook für die Studie anmelden, indem Sie auf »Sign Up« klicken.

Um zu erklären, warum solch eine Werbung aus Datenethik- und Datenschutzperspektive ein Problem ist, lassen Sie uns noch einmal kurz den Lebenszyklus der Werbung durchgehen: Die Werbekund:in stellt der Plattform die Anzeige und einige Targeting-Kriterien zur Verfügung. Der Targeting-Algorithmus der Plattform, ein prädiktives Modell, ermittelt, für welche Individuen die Anzeige nach den genannten Kriterien besonders relevant ist, und ihnen wird die Anzeige angezeigt. Nun muss man jedoch Folgendes wissen: Digitale Werbeprovider, darunter Facebook, Apple und Google, zeigen den

33 Vgl. MD Connect 2017; Denecke u. a. 2015; Guthrie u. a. 2019; Wisk u. a. 2019 und zur kritischen Diskussion Mühlhoff und Willem 2022.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Nutzer:innen nicht bloß individuell zugeschnittene Werbung an, sondern sie zeichnen auch das *engagement* der Nutzer:innen mit der Werbung auf. Es wird beispielsweise vermessen, ob die Nutzer:in sich die Werbung anschaut (wie lange sie beim Scrollen dort stehen bleibt), ob sie auf die Werbung klickt oder sie vielleicht sogar »teilt« (also jemand anderem schickt). Auf die hier abgebildete Werbung konnten Sie sogar klicken, um sich direkt für die Studie anzumelden. Das ist ein Klick, der auf Facebook stattfindet und somit ebenfalls von Facebook registriert wird.

Was passiert nun mit diesen Engagement-Daten? Diese Daten können insbesondere als Trainingsdaten wieder in das prädiktive Modell, das für die Targeting-Entscheidung verantwortlich ist, eingespeist werden. Es handelt sich bei den Engagement-Daten schließlich um wertvolles Feedback darüber, welche Nutzer:innen auf die Anzeige geklickt haben und welche nicht. Damit kann der prädiktive Algorithmus präzisiert werden, um noch treffender vorherzusagen, welche Nutzer:innen in Zukunft auf die Werbung klicken werden und welche nicht. Das prädiktive System wird auf diese Weise immer besser und treffsicherer. Genau dieser Effekt wurde in wissenschaftlichen Publikationen beschrieben. Zum Beispiel wollten Wissenschaftler:innen aus den USA über Facebook-Anzeigen Cannabis-Konsument:innen für eine Untersuchung zu Cannabis-Konsumverhalten akquirieren.[34] Sie beschreiben in ihrer Studie, wie die Targeting-Algorithmen für diese Anzeige zu Beginn relativ schlecht waren, in den ersten Tagen jedoch eine steile »Lernkurve« zeigten. Das Targeting wurde erst nach einigen Tagen

34 Borodovsky u. a. 2018.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Abb. 12: Facebook-Werbung für eine Studie zu Diabetes Typ 2 der Agentur Trialfacts, San Diego, USA. Quelle: Trialfacts, https://trialfacts.com/case-study/effective-clinical-trial-recruitment-plan-narrowing-field-from-500-to-24/.

Tagen sehr effizient.[35] Das ist plausibel, wenn der Algorithmus anhand der gewonnenen Engagement-Daten der bisherigen Nutzer:innen, die die Werbung gezeigt bekommen haben, ständig lernt, seine Vorhersage zu verbessern. Denn kaum eine Nutzer:in gibt auf Facebook explizit an, Cannabis zu konsumieren. Das prädiktive Modell kann diese unbekannten Informationen jedoch nach kurzer Zeit interpolieren – wegen des Feedbacks

35 Ebd., S. 2.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

auf die von den Wissenschaftler:innen geschaltete Werbeanzeige.

Während also eine Werbekampagne auf einer Social-Media-Plattform läuft, wird »in Echtzeit« anhand der Klicks der Nutzer:innen auf diese Anzeige ein prädiktives Modell trainiert, welches mit der Zeit immer genauer vorherzusagen lernt, wer auf die Anzeige klicken wird. Dieses prädiktive Modell gehört der Social-Media-Plattform, denn es handelt sich um proprietäres Datenmaterial dieses Unternehmens. Was ist das für ein Modell im Fall unserer D2M-Werbung? – Das ist natürlich ein Modell, welches mit medizinischen Informationen korreliert. Es ist ein Modell, welches über beliebige Facebook-Nutzer:innen voraussagen kann, ob sie wohl an Diabetes Typ 2 leiden – genauso wie die Plattform im Fall der Studie von Borodovsky et al. ein Modell für Cannabis-Konsum trainieren konnte. Denn diejenigen Nutzer:innen, die auf die Werbung klicken, sind mit hoher Wahrscheinlichkeit selbst von dem Thema betroffen. Über das *engagement* mit der Anzeige verraten die Nutzer:innen unwissentlich höchst sensible Informationen, die als Trainingsdaten in das prädiktive Modell eingespeist werden. So wird die Social-Media-Plattform mit der impliziten Unterstützung der externen Forschungseinrichtung, welche die Werbung für eine klinische Studie beauftragt, dazu in die Lage versetzt, ein prädiktives Modell zu trainieren, welches über *beliebige* Facebook-Nutzer:innen – heute oder in der Zukunft – abschätzen kann, ob sie wahrscheinlich an Diabetes Typ 2 leiden.[36]

Dieses Schema benennt ein umfassendes und gravierendes, zugleich jedoch weithin vernachlässigtes Daten-

36 Mühlhoff und Willem 2022.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

ethik- und Datenschutzproblem im Kontext individualisierter Werbung. Es ist davon auszugehen, dass die Provider digitaler Werbung für jedes *beliebige* Thema, zu dem Werbung geschaltet wird, ein prädiktives Modell trainieren, welches dann auf beliebige dritte Nutzer:innen angewandt werden kann, um den Bezug der Nutzer:in auf den Inhalt der Werbung abzuschätzen. Stellen Sie sich vor, bei welchen Themen allein im Medizinbereich dies höchst brisant ist. Denken Sie an Werbung für Alkohol- und Drogenentzugstherapie, HIV-Prävention, Hepatitis-B-Therapie, Therapie für psychologische Krankheiten wie Depression, Angststörungen, Paranoia etc.

Zu bedenken ist dabei, das die Engagement-Daten und die damit trainierten prädiktiven Modelle den Plattformen rechtlich zugeordnet sind. Die medizinischen Forschungseinrichtungen gelangen nicht an diese Daten, auch wenn diese für ihre Forschung mitunter extrem nützlich sein könnten. Tatsächlich bezahlen die Forschungseinrichtungen die Werbeplattform sogar dafür, das Targeting für sie zu übernehmen.

Was würden Sie mit den Daten und den daraus trainierten prädiktiven Modellen tun, wenn Sie eine Social-Media-Plattform wären? Sie würden sie natürlich verkaufen (genau genommen würden Sie Zugriff auf die prädiktiven Modelle verkaufen, nicht die Daten selbst). Sie könnten sich nämlich noch ganz andere Kund:innen vorstellen als die Betreiber klinischer Studien, die an diesen Prädiktionen Interesse hätten. Denken Sie an die Versicherungsindustrie, an die automatische Klassifikation von Bewerber:innen bei Einstellungsverfahren, an Credit Scoring. In diesen Sektoren gibt es zahlungskräftige Abnehmer solcher aus Engagement-Daten hergestellter Vorhersageprodukte. Die Kund:innen aus diesen

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Sektoren werden natürlich nicht am Frontend der Plattform sichtbar, sondern am Backend bedient, sodass wir sie als normaler Nutzer:innen nie sehen.

Die Beispiele dieses Kapitels lassen ein wichtiges Fazit zu. Prädiktive Modelle – als eine der häufigsten Anwendungsformen des maschinellen Lernens heute – sind zunächst Human-Aided AI-Systeme. Denn nur der fortgesetzte Dateninput von Nutzer:innen (indem diese zum Beispiel auf Werbung klicken oder nicht) ermöglicht die Herstellung solcher Modelle. Doch zugleich haben solche KI-Systeme, zu denen wir alle beitragen durch unsere Klicks im Internet, eine Rückwirkung auf uns. Denn diese Systeme beeinflussen, welche Informationen wir angezeigt bekommen, zu welchen Ressourcen (Kredite, Versicherungen, Bildungsangebote, Jobs) wir Zugang erhalten, gegebenenfalls sogar, ob Sicherheitsbehörden oder Jugendschutz intervenieren. Prädiktive Anwendungen von KI strukturieren das gesellschaftliche Feld, den Zugang zu Chancen, Möglichkeiten, Ressourcen und Informationen in unserer Gesellschaft. Es handelt sich bei diesen Systemen um großskalige, von den Datenunternehmen betriebene kybernetische Feedback-Schleifen:[37] Wir tragen mit unseren Daten zu den Systemen bei; die Systeme modulieren wiederum unsere Bewegungsmöglichkeiten, unser Wissen und unsere Gefühle.

Hierbei gibt es zahlreiche ethische Probleme. Man könnte jetzt auf Biases in den Vorhersagen und Diskriminierung eingehen.[38] Man könnte auch auf das Problem der (fehlenden) Erklärbarkeit und Transparenz solcher

37 Vgl. Nosthoff und Maschewski 2019.

38 Siehe ausführlicher Buolamwini und Gebru 2018; Coeckelbergh 2020; O'Neil 2016; Mittelstadt u. a. 2016.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Verfahren, auf Fragen der Accountability und moralischen Verantwortlichkeit eingehen. Ich werde diese wichtigen Themen für heute Abend gezielt aussparen. Stattdessen möchte ich ein weiteres, meines Erachtens fundamentales ethisches Problem in diesem Zusammenhang herausgreifen und näher ausführen: nämlich die Auswirkungen von prädiktiver Analytik auf Privatheit.

Teil 3: Kollektiver Datenschutz und prädiktive Privatheit

Für eine Besprechung aus der Perspektive von Datenschutz und Datenethik müssen wir uns noch einmal zwei zentrale Strukturmerkmale prädiktiver Analytik vor Augen führen, auf die ich in den folgenden beiden Unterabschnitten jeweils etwas ausführlicher eingehen werde. Erstens ist prädiktive Analytik von einer *Eskalation der Sensibilität* der Daten gekennzeichnet: Aus leicht zugänglichen Daten über ein Individuum können mittels prädiktiver Modelle schwer zugängliche Informationen abgeschätzt werden. Zweitens besitzt prädiktive Analytik eine *kollektive Verursachungsstruktur*: Die Datenfreigiebigkeit einer hinreichend großen Menge Nutzer:innen ist dafür verantwortlich, dass prädiktive Modelle trainiert werden können, die dann auf beliebige *andere* Nutzer:innen zur Abschätzung von Informationen über sie angewendet werden können.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Es ist, wie beschrieben, der prädiktiven Analytik eigentümlich, dass sie aus Informationen, die die meisten Nutzer:innen für relativ wenig sensibel halten, wie beispielsweise ihre Facebook-Likes oder ihren Browserverlauf, subjektiv betrachtet viel sensiblere Informationen ableiten kann. Prädiktive Analytik ermöglicht eine *Eskalation der Sensibilität* der Daten. Sie denken, Sie haben doch nur Facebook-Likes, Ihren GPS-Standort oder Ihre Kaufhistorie auf Amazon preisgegeben, aber in Wirklichkeit leiten die Plattformunternehmen daraus Schätzung noch weiterer Informationen über Sie ab, etwa Ihre sexuelle Orientierung, ob Sie Substanzen missbrauchen, psychisch labil oder schwanger sind. Solche Vorhersagen können übrigens auch Informationen umfassen, die Sie selbst noch gar nicht kennen – denken Sie an Krankheitsprognosen, von denen Sie vielleicht nichts wissen oder nichts wissen wollen; oder denken Sie an ein Scoring Ihrer Bonität oder Ihres Kreditausfallrisikos, das sind Zahlen, die Sie normalerweise nicht zu Gesicht bekommen.

Entlang dieses ersten Strukturmerkmals von prädiktiver Analytik müssen wir also festhalten, dass prädiktive Analytik potenziell in die Privatsphäre beliebiger Menschen eingreift – also in Ihr Recht, zu wissen und zu kontrollieren, welche Informationen über Sie von einer anderen Instanz verarbeitet werden. In meiner Forschung argumentiere ich, dass prädiktive Analytik eine neue Form der Verletzung von Privatsphäre konstituiert. Ein Begriffsvorschlag, den ich in den letzten Jahren erarbeitet habe, ist »prädiktive Privatheit« oder »predictive

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

privacy«.[39] Unter diesem Titel führe ich eine ethische Wertedebatte, in der es darum geht, dass wir unser Verständnis von Privatheit angesichts von Big Data und KI auf ein Bewusstsein um die Fähigkeiten prädiktiver Analytik ausweiten müssen. Wir benötigen ein Verständnis von Privatheit, welches umfasst, dass die Privatheit einer Person oder Gruppe auch dann verletzt ist, wenn sensible Informationen ohne ihr Wissen und gegen ihren Willen über sie *vorhergesagt* werden.[40] Ihre Privatheit ist nicht nur dadurch verletzbar, dass Informationen, die Sie irgendwo *preisgegeben* haben (zum Beispiel bei Ihrer Ärzt:in oder in einem Onlineshop) missbräuchlich weitergegeben oder von Hacker:innen entwendet werden. Sondern, so der normative Beitrag, Ihre Privatheit ist auch dann verletzt, wenn Informationen über Sie *abgeleitet* werden, die Sie niemals irgendwo angegeben haben oder vielleicht sogar selbst gar nicht kennen.

Mit dem Begriff der prädiktiven Privatheit findet somit eine Bereichserweiterung statt: In die Domäne ihrer informationellen Selbstbestimmung fallen nicht nur Informationen, die Sie kontrollieren *können*, sondern auch solche, die über Sie geschätzt werden, unabhängig vom Wahrheitsgehalt der Schätzung. Das so zu sehen, ist keine Selbstverständlichkeit, denn den meisten Menschen ist aktuell nicht klar, dass digitale Geschäftsmodelle heute systematisch auf der Verwendung abgeleiteter Informationen über (ggf. sogar anonyme) Datensubjekte beruhen. Übrigens ist der Tatbestand der prädiktiven Verletzung von Privatheit nicht davon abhängig, ob die über Sie vorhergesagten Informationen korrekt sind oder

39 Die folgende Darstellung beruht auf Mühlhoff 2021, 2022.
40 Ebd.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

nicht. Es geht darum, *dass* etwas vorhergesagt wird, das Sie nicht preisgeben möchten oder können, insofern diese Informationen dann in eine Entscheidung über Sie oder eine individualisierte Behandlung einfließen. Auch falsche Vorhersagen greifen dann in Ihr Recht auf informationelle Selbstbestimmung ein.

Die erste Herausforderung, die mit dem Begriff der prädiktiven Privatheit verbunden ist, besteht darin, für diese neue Form der Verletzung von Privatheit ein allgemeines Bewusstsein zu schaffen. Darauf aufbauend benötigen wir eine gesellschaftliche Debatte, wie sich der ethische Wert der prädiktiven Privatheit zu anderen Werten verhält, gegen die er in konkreten Situationen gegebenenfalls abgewogen werden muss. Zum Beispiel ist es naheliegend, dass viele Patient:innen der Verwendung neuer medizinischer Diagnoseverfahren auf der Grundlage von prädiktiver Analytik zustimmen (und damit situativ auf Wahrung ihrer prädiktiven Privatsphäre zugunsten des Wertes ihrer Gesundheit verzichten).

Schwieriger und kontroverser ist zum Beispiel die Abwägung prädiktiver Privatheit gegen den Wert der öffentlichen Sicherheit. Wollen wir, dass Menschen aufgrund von Prognosen über ihr Verhalten präventiv festgesetzt werden? In zahlreichen US-Bundesstaaten werden zum Beispiel im Justizsystem prädiktive Modelle für die kriminelle Rückfallwahrscheinlichkeit von Straftäter:innen verwendet, etwa wenn es um Entscheidungen über Hafterleichterung oder das Aussetzen von Haftstrafen auf Bewährung geht. Es hat sich herausgestellt, dass solche Modelle gravierende rassistische Biases aufweisen und bestehende Diskriminierungsmuster der US-amerikanischen Gesellschaft im Gewand einer datenbasierten und daher vermeintlich objektiven Beurteilung fort-

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

schreiben.[41] Doch nehmen wir einmal an – was allerdings unrealistisch und daher ein reines Gedankenexperiment ist –, es gäbe ein Vorhersagemodell für kriminelle Rückfallwahrscheinlichkeit, das keine Biases besitzt (ich denke nicht, dass es das jemals geben könnte) – selbst dann wäre es ethisch noch längst nicht klar, dass wir ein solches Vorhersagemodell als Grundlage für Entscheidungen über Hafterleichterung verwenden dürfen. Denn es ist höchst streitbar, ob wir Menschen aufgrund von *Prognosen* über ihr zukünftiges Verhalten – ein Verhalten, das sie selbst beeinflussen können – zum Beispiel in ihrer Freiheit einschränken dürfen. Dieses allgemeinere Problem einer präventiven Inhaftierung bekommt im Kontext prädiktiver Analytik eine neue Schärfe, da diese Prognosen nicht aus umfassenden, von Menschen und im Einzelfall durchgeführten Begutachtungen resultieren, sondern maschinell erstellt werden, anhand von statistischen Modellen, die nicht nach Gründen fragen und nicht empathiefähig sind, sondern lediglich einen Abgleich des konkreten Falls mit einigen Tausend anderen Fällen durchführen.

Während ich diese ethische Frage in dem vorliegenden Rahmen nicht abschließend behandeln kann, da sie vom eigentlichen Anliegen wegführt, möchte ich mit dieser Diskussion nur Folgendes zeigen: Mit dem Begriff der prädiktiven Privatheit wird ein grundlegendes ethisches Problem prädiktiver Analytik sichtbar, das noch fundamentaler ist als das viel besprochene Problem von Biases. Selbst wenn ein prädiktives Modell *keine* unfairen Biases aufweisen würde, wäre ethisch noch nicht geklärt, ob wir es für automatisierte Entscheidungen über Men-

41 Angwin u. a. 2016.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

schen verwenden dürfen. Die zurzeit prominente Debatte zu Biases in automatisierten Entscheidungsverfahren ist enorm wichtig und von unmittelbarer sozialer Relevanz, da sie auf reale Mechanismen der Diskriminierung und sozialen Ungleichheit hinweist. Sie darf aber nicht dahingehend missverstanden werden, dass ein Vorhersageverfahren in dem Moment ethisch und politisch unbedenklich wird, wo es keine Biases mehr aufweist. Eine Ethik der KI darf sich nicht in den Dienst einer bloßen »Korrektur« solcher Modelle durch das Ausmerzen von Biases stellen, sondern muss die Verwendung von KI-Modellen für bestimmte Entscheidungsverfahren grundsätzlich hinterfragen.

Ein grundsätzliches ethisches Problem der Verwendung prädiktiver Modelle für automatisiertes Entscheiden ist zum Beispiel das Problem der *prediction gap*:[42] Wenn eine Entscheidung über eine Person an einer Prognose über sie orientiert wird, dann muss aus der Liste möglicher Werte der prognostizierten Variablen *eine* Option ausgewählt werden, entsprechend derer die Person dann real behandelt wird. Ein prädiktives Modell trifft selten eine eindeutige Entscheidung, sondern der Output ist eine Liste möglicher Werte der Zielvariable, versehen mit Wahrscheinlichkeitsgewichten. Ein Vorhersagemodell für kriminelle Rückfälligkeit würde in einem konkreten Fall zum Beispiel für »Rückfälligkeit« die Wahrscheinlichkeit 0,65 und für »keine Rückfälligkeit« die Wahrscheinlichkeit 0,35 ausrechnen. Wenn an so einer Prognose eine Entscheidung festgemacht und die Person individualisiert behandelt werden soll, muss aber aus der Liste der *möglichen* Werte der Zielvariable ein *kon-*

42 Ich folge hier der Studie Mühlhoff 2021.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

kreter Wert ausgewählt werden – man kann die Person ja nicht gleichzeitig freilassen und weiter einsperren. Es liegt nahe, den Wert mit dem höchsten Wahrscheinlichkeitsgewicht zu wählen. In diesem Moment überschreiten wir das, was ich als *prediction gap* bezeichne: Wir legen die Person auf ihr *wahrscheinlichstes* Verhalten fest – und verweigern es ihr, ein »Ausreißer« aus dieser statistischen Analyse zu sein. Oder ethisch gesprochen verweigern wir ihr Autonomie, Handlungsfreiheit und eine prinzipielle Diversität der menschlichen Erscheinungsformen. Im Fall krimineller Rückfälligkeit gilt grundsätzlich: Die Person *könnte* wieder eine Straftat begehen, sie *könnte* es aber auch sein lassen, und da es sich um einen Menschen handelt, ist ihr tatsächlich beides zuzutrauen. Davon geht auch unsere Rechtsordnung aus, die Straffälligkeit an Willensentscheidungen anknüpft und zum Beispiel für einen Rücktritt von einer versuchten Straftat Straferleichterungen vorsieht. Wir beschreiten einen logisch, epistemologisch und ethisch zweifelhaften Weg, wenn wir eine Wahrscheinlichkeitsverteilung in unzulässiger Weise auf eine Punkt-Vorhersage vereindeutigen. Dieser Schritt bedeutet im Sinne der Bayes'schen Statistik, dass wir eine rational informierte Wette über das betroffene Individuum eingehen und es so behandeln »als ob« wir sicher wüssten, dass es rückfällig werden wird.[43] Dieses Prinzip des Wettens führt dazu, dass wir das konkrete Individuum auf ein Stereotyp festsetzen und mit anderen Individuen in Gruppenhaft nehmen, die bei gleicher Datenlage das »typische« Verhalten zeigen.

43 Vgl. zum Aspekt eines Bayes'schen Verständnisses von Wahrscheinlichkeit auch Joque 2022.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Die realen Auswirkungen prädiktiver Analytik sind nicht auf die individuellen Konsequenzen begrenzt. Daher werde ich die Individuen-bezogene Perspektive dieses ersten Unterabschnitts nun durch eine zweite, kollektivistische Perspektive ergänzen, die von dem zweiten eingangs genannten Strukturmerkmal prädiktiver Analytik ausgeht.

B. Kollektive Verantwortung

Das zweite strukturelle Merkmal prädiktiver Analytik ist noch brisanter, jedoch zugleich schwieriger nachzuvollziehen: Wenn ein prädiktives Modell sensible Informationen über Sie abschätzt, dann erfolgt das anhand der Daten *vieler anderer* Individuen. Prädiktive Analysen werden mit den – potenziell anonymisierten – Daten von vielen Millionen Individuen trainiert und somit erst durch unsere kollektive Datenerzeugungspraxis ermöglicht. Das Individuum, auf das ein prädiktives Modell angewendet wird, muss selbst nicht in dem Trainingsdatensatz enthalten sein, mit dem das Modell erstellt wurde. Das heißt, wir alle tragen bei der Verwendung vernetzter digitaler Medien dazu bei, dass zum Beispiel Plattformunternehmen Vorhersagemodelle bauen können, mit denen wiederum *andere* Individuen – nicht notwendigerweise wir selbst – unterschiedlich behandelt und diskriminiert werden können.

Das ist eine wichtige Botschaft vor allem an all jene, die denken, sie selbst »haben doch nichts zu verbergen« und können deshalb Google oder Dropbox, Facebook, Instagram oder Apple ruhig ihre Daten überlassen. Es braucht die Daten der zahlreichen vermeintlich »normalen« oder sich selbst für normal haltenden Nutzer:innen,

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

um prädiktive Modelle zu trainieren, die dann *andere* Individuen als mutmaßlich gefährlich, krank, riskant etc. klassifizieren. Unsere Datenfreigiebigkeit ist also ein für das Gemeinwesen relevantes Verhalten, gerade auch dann, wenn wir keine negativen Auswirkungen für uns selbst befürchten oder denken, »das Unternehmen weiß doch eh schon alles über mich«. *Unsere* Daten ermöglichen automatisierte Wetten auf das zukünftige Verhalten oder unbekannte Eigenschaften von beliebigen anderen Menschen und tragen auf diese Weise zur Stabilisierung und Steigerung sozialer Ungleichheit und ökonomischer Ausbeutungsmuster von gesamtgesellschaftlichem Ausmaß bei. Denn prädiktive Analytik wird oft dazu verwendet, ohnehin schlechter gestellten Menschen schlechtere Konditionen anzubieten und ihnen den Zugang zu Jobs, Bildung und wohlfahrtsstaatlichen Ressourcen zu erschweren.[44]

Weil unsere Daten potenziell anderen schaden, besitzt prädiktive Analytik eine *kollektive Verursachungsstruktur*. Diese KI-Systeme sind nur möglich, weil viele von uns als Nutzer:innen seriell die gleichen Entscheidungen treffen – nämlich, dass wir Daten täglich preisgeben, die wir subjektiv für ausreichend harmlos halten, sodass für uns individuell betrachtet der Nutzen bestimmter digital vernetzter Dienste gegenüber den möglichen Risiken überwiegt. Autonomie, Gleichbehandlung und Privatheit beliebiger Personen werden durch prädiktive Analytik also genau deshalb verletzbar, weil für hinreichend viele *andere* Menschen die negativen Effekte ihrer Datenpraxis nicht ausreichend spürbar sind oder nicht ausreichend reflektiert werden. Oft reicht eine gesellschaftliche Min-

44 O'Neil 2016; Eubanks 2017; Noble 2018.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

derheit, denn im Kontext sozialer Medien genügt es, wenn einige wenige Prozent der Nutzer:innen bestimmte Daten preisgeben. Es entsteht dann trotzdem eine ausreichend große Mengen Datenpunkte, um damit prädiktive Modelle zu trainieren.

Eine Eselsbrücke, um sich die kollektive Dimension des eigenen Datenverhaltens besser vorzustellen, ist die Analogie zur Umweltverschmutzung durch Autoabgase. Wir können uns die Daten, die bei unserer täglichen Benutzung vernetzter digitaler Dienste anfallen, wie eine »Datenverschmutzung« vorstellen.[45] Wenn Sie mit dem Auto fahren, schaden Sie offensichtlich nicht nur sich selbst, sondern auch anderen Menschen, insofern Sie zur Minderung der Luftqualität und zum Klimawandel beitragen. Genauso schaden auch die Daten, die wir preisgeben, nicht nur uns selbst, sondern der Gesellschaft im Ganzen. Ökonomisch gesprochen könnte man sagen, Daten sind mit sozialen Externalitäten verbunden. Das sind Folgekosten unserer Datenpraktiken, die aber nicht notwendig bei uns selbst anfallen und somit nicht in unsere individuelle Kosten-Nutzen-Abrechnung eingepreist sind, sondern von anderen oder vom Gemeinwesen getragen werden müssen.

Diese Analogie zur Umweltverschmutzung ist eine griffige Metapher und kommunikationsstrategisch nützlich, hat jedoch im Detail ihre Grenzen. Erstens ist es unstrittig, dass Abgase rein negative Externalitäten darstellen, während Datenaggregation auch positive Effekte haben kann. Denn es gibt Anwendungen prädiktiver Modelle, die nützlich für unsere Gesellschaft sind, zum Beispiel Pandemien vorhersagen oder die medizinische

45 Ben-Shahar 2019.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Diagnostik verbessern. Sie haben also bei der Thematisierung von Daten als soziale Externalität das Problem, erst einmal eine Debatte über gesellschaftlich wünschenswerte und nicht wünschenswerte Effekte von prädiktiver Analytik führen zu müssen, bevor Sie wissen, in welchem Fall das Vorzeichen der Externalität positiv oder negativ ist. Zweitens bricht die Analogie auch in Bezug darauf, wie sehr die Externalität mit der Anzahl der Mitverursacher:innen skaliert. Fahren nur noch halb so viele Menschen Auto, ist grob gesagt die Luftqualität doppelt so gut. Das Gleiche gilt in Bezug auf die Datensammlung in sozialen Medien jedoch nicht: Halb so viele Datenpunkte sind oft immer noch genug, um ein prädiktives Modell zu trainieren, das auf alle anwendbar ist. Und wenn nicht, werden die Modelle trotzdem trainiert und sind dann eben lediglich ungenauer.

Insgesamt zeigt sich, dass die kollektive Verursachungsstruktur prädiktiver Analytik in der Ethik eine relativ schwer zu fassende Konstellation darstellt. Klassische Ethik fokussiert auf das moralische Verhalten des einzelnen Akteurs. Niemand Einzelnes von uns ist jedoch moralisch dafür verantwortlich, dass Plattformunternehmen prädiktive Modelle herstellen können. Es liegt hier auch kein Phänomen »distribuierter Verantwortung« vor, denn das würde suggerieren, dass jede:r von uns einen kleinen, jedoch kumulativen Teil der Gesamtverantwortung trägt. Auch das ist nicht plausibel. Vielmehr handelt es sich um ein kollektiv verursachtes Phänomen, wobei allerdings das Kollektiv der Nutzer:innen eigentlich gar kein *Kollektiv* im starken Sinne einer gemeinsamen Identität, eines geteilten Bewusstseins oder einer gemeinsamen Zielorientierung ist, sondern eher ein Zustand unreflektierter und »serieller Kollektivität« in dem

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Sinne, wie die Philosophin Iris Marion Young es im Anschluss an Jean-Paul Sartre mit Blick auf Gender-Strukturen formuliert hat.[46] Serielle Kollektivität bedeutet hier: Indem hinreichend viele von uns sich ähnlich verhalten – also ohne Hinterfragen oder gegebenenfalls nach individueller Kosten-Nutzen-Abwägung vernetzte digitale Dienste benutzen –, entsteht eine strukturelle Konfiguration, eine weder komplett zufällige noch komplett intentionale Parallelität von Praktiken, Reflexions- und Verhaltensweisen, die von gesamtgesellschaftlichem Ausmaß ist. Aufgrund der Verwobenheit dieser Praktiken mit der Ermöglichung prädiktiver Analytik gewinnt diese Parallelität der Praktiken und Verhaltensweisen im Ganzen die Qualität einer sozialen Struktur, und auf diese Struktur muss sich eine kritische Ethik der KI beziehen.[47] Es ist eine *strukturbezogene Ethik der KI* zu formulieren, oder genauer gesagt: eine Ethik emergenter Strukturen – so würde ich das Projekt meiner Forschung in Osnabrück der nächsten Jahre bezeichnen.

Auch für die Art und Weise, wie Privacy und Datenschutz weitläufig verstanden werden, ist die kollektive Verursachungsstruktur prädiktiver Analytik etwas Neues. Denn Datenschutz wird im öffentlichen Diskurs meist individualistisch geframed, im Sinne von »jede:r hat das Verfügungsrecht über seine eigenen Daten«. Es wird (fälschlicherweise) suggeriert, dass dem Datenschutz Genüge getan sei, wenn man vor der Verarbeitung der eigenen Daten eine Einwilligung erteilt oder wenn Daten nur anonymisiert erhoben werden. Das ist die liberalistische, individualistische Auslegung von Datenschutz, die

46 Young 1994.
47 Siehe ausführlicher Mühlhoff 2020b.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

das ethische Problem genauso auf die Individuen abschiebt und privatisiert, wie die Individualethik es tut. Ein Datenschutz, der auf die Sorge um die individuelle Kontrolle über Informationspreisgabe verengt wird, ist allerdings gegenüber der seriell-kollektiven Verursachungsstruktur prädiktiver Analytik zahnlos.

Diese Unzulänglichkeit des Datenschutzes bezieht sich auch auf den gegenwärtigen Regulierungsstand im Kontext der europäischen Datenschutzgrundverordnung (DSGVO) und der nationalen Datenschutzgesetze. Ich möchte nur zwei Gründe dafür ansprechen:[48] Erstens genügen zum Training prädiktiver Modelle anonyme Daten – es müssen, um mit Abbildung 8 zu sprechen, nur die grünen und roten Datenpaare vorhanden sein, ein konkreter Personenbezug ist dabei nicht wichtig und kann mit geeigneten Methoden entfernt werden. Anonyme Daten fallen jedoch nicht in den Regelungsbereich der DSGVO. Erfolgt die Erstellung prädiktiver Modelle also mittels anonymer Daten, ist die DSGVO gar nicht anwendbar, da diese nur personenbezogene Daten erfasst. Mit dem Versprechen einer anonymen Verarbeitung sind überdies auch sensible Daten vergleichsweise leicht von Internetnutzer:innen erhebbar. Des Weiteren fallen nach einer Anonymisierung nicht nur die Trainingsdaten der prädiktiven Modelle aus dem Anwendungsbereich der DSGVO, auch die trainierten Modelle sind dann erfasst: Es handelt sich bei einem prädiktiven Modell (betrachtet als Matrix von Gewichten und Parametern) selbst um hochgradig aggregierte anonymisierte

48 Siehe ausführlich Mühlhoff und Ruschemeier 2022; Ruschemeier 2021.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Daten. Diese können daher unreguliert verarbeitet und sogar verkauft werden.

Zweitens ist die praktisch wichtigste Rechtsgrundlage der Einwilligung in der DSGVO ein Schwachpunkt, der zahlreiche Geschäftsmodelle ermöglicht, die prädiktive Analytik verwenden. Denn eine Einwilligung für die (anonymisierte oder nicht anonymisierte) Verwendung von Social-Media-Daten ist von Nutzer:innen relativ leicht zu erhalten, auch wenn sie in den meisten Fällen in digitalen Kontexten weder tatsächlich freiwillig noch informiert erfolgt. Das Prinzip der Einwilligung bildet zudem die kollektive Verursachungsstruktur prädiktiver Modelle nicht ab. Es wird ausgeblendet, dass das Datensubjekt mit seiner Einwilligung eine Entscheidung für viele andere Individuen und ultimativ die Gesellschaft im Ganzen trifft. Doch wie wäre dieses Problem in liberalen Rechtssystemen zu lösen, ohne den Menschen die Verwendung vernetzter digitaler Dienste zu verbieten? Das führt zu der Forschungsfrage, wie man eigentlich einen kollektiven Datenschutz oder einen kollektiven Grundrechtsschutz angesichts von KI-Technologie formulieren kann.[49]

Schluss: Die Macht der Daten

Von den Human-Aided AI-Apparaten allgemein bis zur prädiktiven Analytik als spezifischer Anwendungsdomäne: Eine Struktur ist deutlich erkennbar, nämlich, dass wir alle mit drin hängen in dem, was KI-Technologie heute ist und bewirkt. Auf die Frage hin, was man tun

49 Mühlhoff und Ruschemeier 2022.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

könne, um sich selbst, andere oder die Gesellschaft im Ganzen vor den gravierenden Risiken dieser Technologie zu schützen, haben wir schon gesehen, dass Technikverweigerung oder »Google löschen« kaum vielversprechende Wege sind, solang wir es nicht alle tun. Und selbst das wäre gar nicht unbedingt wünschenswert, weil wir dann die enormen gesellschaftlichen Vorteile und Fortschritte durch diese Dienste und Technologien verpassen würden.

Für eine Ethik der künstlichen Intelligenz bedeutet diese Situation, dass wir eine Debattenebene jenseits der klassischen ethischen Frage nach dem richtigen Handeln des moralisch verantwortlichen Einzelakteurs erschließen müssen. Was jedoch sind die Angriffspunkte und Einflussmöglichkeiten ethischer Besprechung, wenn es schlicht keinen großen Unterschied macht, ob Einzelne von uns das Spiel mitmachen oder nicht?

Im Kontext dieses Vortrags ging es mir zunächst darum, das gesellschaftliche Bewusstsein für die Funktionsweise von datenbasierter KI und prädiktiver Analytik zu fördern. Im Fokus stand dabei die Doppelstruktur unserer unwillkürlichen Mitwirkung an und unausweichlichen Betroffenheit von diesen Systemen. Wir alle haben in unserem Leben mit den Auswirkungen dieser KI-Systeme zu tun – bewusst oder unbewusst – und die meisten von uns tragen auch selbst mit ihren Daten zum Betrieb dieser Systeme bei. Und doch bedeutet Verantwortungsübernahme seitens der Nutzer:innen nicht unbedingt, vollständig auf die Verwendung dieser Geräte und Dienste zu verzichten. Denn erstens ist zweifelhaft, ob das überhaupt möglich ist und im Einzelfall zielführend wäre; zweitens kommt es vielmehr darauf an, die Technologien reflektiert und mit Einblick in ihre Funkti-

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

onsweise zu benutzen, anstatt sie abzulehnen. Und drittens schließlich haben wir noch einen weiteren gewichtigen Hebel gegen die Datenindustrie in der Hand, dessen Möglichkeiten noch nicht voll ausgeschöpft sind: bessere Regulierung. Die Daten- und KI-Industrie hat zurzeit vergleichsweise geringe regulatorische Auflagen in Bezug auf die Sammlung und Verwendung insbesondere anonymisierter Daten. Diese Industrie *könnte* viel besser reguliert werden. Es *könnte* an stärkere Bedingungen geknüpft werden, wann und ob ein Unternehmen die Daten vieler Einzelpersonen zusammenfassen und daraus ein prädiktives Modell trainieren, oder wann und ob ein Unternehmen die kostenlose Arbeitsleistung vieler Nutzer:innen zu einem großen distribuierten Rechennetzwerk zusammensetzen darf.

Was wir auf dem Weg zu einer besseren Regulierung prädiktiver Analytik benötigen, ist das Bewusstsein dafür, dass im Kontext von KI-Technologie die Sensibilität aggregierter Daten größer ist als die Summe der von uns jeweils einzeln wahrgenommenen Sensibilität der isolierten Datenpunkte (z. B. der eigenen Facebook-Likes). Wer die Daten vieler Individuen hat, interessiert sich nicht mehr für das Einzelindividuum – es geht nicht um Spionage oder Stalking in Bezug auf *Sie* im Einzelnen. Das Ziel prädiktiver Analytik ist das automatisierte Sortieren und die algorithmische Verwaltung großer Menschenmengen auf der Grundlage von Mustern und Regularitäten in den Daten.[50] Der Ansatzpunkt für eine Regulierung dieser Technologie sollte sein, dass es jedoch *nicht selbstverständlich* ist, dass ein Unternehmen, welches uns Informations- oder Kommunikationsdienstleis-

50 Mühlhoff 2020b.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

tungen anbietet, im Gegenzug das enorme Informationspotenzial der dabei anfallenden Massendaten ausbeuten und gegen uns verwenden darf.

Für eine treffende ethische, datenschützerische und regulatorische Besprechung von datenbasierten KI-Systemen ist es darum erforderlich, das ganze Bild zu sehen: Das große Problem ist nicht die Verletzung der Privatheit eines *Einzelnen* von uns, sondern die Herstellung prädiktiver Modelle, die jederzeit und automatisch auf viele Menschen angewendet werden können. Es geht um die *potenziellen* Schäden oder Auswirkungen prädiktiver Modelle – um die »Streueffekte«, wenn man so möchte. Um es in den drastischen Worten der Mathematikerin Cathy O'Neil zu sagen: Ein prädiktives Modell ist wie eine »mathematische Massenvernichtungswaffe« in den Händen eines Privatunternehmens.[51] Genauso wie wir Waffenbesitz verbieten und nicht erst die Verwendung der Waffe, sollten wir bereits die Herstellung und Verbreitung prädiktive Modelle regulieren.

Ethik der KI muss über Macht sprechen

Etwas weniger metaphorisch ausgedrückt ist das Anliegen einer ethischen Debatte und einer wirkungsvollen Regulierung von datenbasierter KI-Technologie eine Beschränkung der *Macht*, die sich in den Händen der Operateure von KI-Systemen sammelt. Wenn das individuelle Handeln kein fruchtbarer Ansatzpunkt für ethische Besprechung ist und die individuellen Auswirkungen von KI-Technologie ein bei Weitem noch nicht vollständiges Bild des Problems ergeben, dann könnte

51 O'Neil 2016.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

tatsächlich eine machtanalytische Perspektive der Schlüssel für eine wirkungsvollere Herangehensweise in der Ethik der KI sein. Im Fall von prädiktiven Modellen spreche ich von »Vorhersagemacht« – *prediction power*.[52] Dies beschreibt das mit dem Besitz prädiktiver Modelle verbundene Vermögen, über beliebige Individuen Vorhersagen zu treffen. Eine zeitgemäße Ethik der KI muss auf einem breiten Verständnis der Macht der KI-Systeme aufbauen, das sich nicht auf die Verwendung von KI für Vorhersagen beschränkt. Diese Macht ist ein komplexes Phänomen, welches sich auf mehreren Skalen und mehreren Ebenen vollzieht. Nur eine dieser Ebenen ist der Pol der (Daten-)Aggregation: Daten konzentrieren sich bei wenigen globalen Akteuren. Wer diese Daten hat, hat eine erhebliche ökonomische und politische Macht, bis hin zur Gestaltungsmacht über Lebenswelten und Lebensstile der Menschen.

Eine zweite Ebene bilden die sozialen Strukturierungseffekte von KI-Systemen. Damit meine ich die Auswirkungen datenbasierter KI in der Form unterschiedlicher Verteilung von Gütern und Ressourcen – zum Beispiel von Wissen, Bildung, Gesundheit, Arbeit, Partizipationsmöglichkeiten, gesellschaftlichen Chancen. Für eine zeitgemäße Ethik der KI benötigen wir einen Begriff der digitalisierten sozialen Ungleichheit und einen Blick für Verteilungsfragen; wir benötigen ein Verständnis dafür, dass KI in ihrer Verkoppelung mit vernetzten digitalen Medien eine Strukturbedingung unseres sozialen, kommunikativen, ökonomischen, politischen, wirtschaftlichen Lebens ist.

52 Mühlhoff und Ruschemeier 2022.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Die dritte Ebene einer Machtanalytik der KI-Systeme betrifft uns selbst, die Subjekte, die durch vernetzte Interfaces in diese Systeme eingebunden werden. Wir müssen analysieren, wie die Benutzeroberflächen eines Handys, eines Online-Spiels oder der sozialen Medien unsere Weltbeziehungen, unser soziales Bewusstsein, unsere Reflexivität prägen. Wie wir über uns selbst oder unsere sozialen Beziehungen nachdenken und wie wir auf Wissen zugreifen, ist hochgradig durch unsere Einbindung in vernetzte digitale Medien geformt. Um das zu analysieren, ist es sinnvoll, einen Begriff der digital-medial produzierten *Subjektivität* in Anschlag zu bringen.[53] Damit wird einerseits erklärbar, wie unser Denken, Fühlen und Handeln durch KI-Systeme geprägt und strukturiert wird und es doch gleichzeitig *unser* Denken, Fühlen und Handeln ist, mit dem wir die KI-System erst möglich machen und ihnen zu ihrer Macht verhelfen. Diese scheinbar zirkuläre Doppelstruktur ist der Kern einer kritischen Philosophie der Subjektivität. Vernetzte digitale Medien sind einerseits darauf angewiesen, dass wir sie benutzen und Inhalte generieren, deshalb machen sie uns abhängig, locken uns in die Falle des endlosen Scrollens oder Klickens und lassen unsere kulturellen Kompetenzen verkümmern, ohne Google Wissen zu recherchieren oder den Weg zu einem anderen Ort zu finden. Zugleich *wollen* wir diese Medien auch benutzen, unsere Prägung *durch* die Medien ist zum Selbstläufer geworden, denn die Benutzung ist lustvoll und effizient, die Präsenz in Onlinemedien ist Teil unserer Identität geworden. Dieser

53 Siehe zu dieser Programmatik ausführlicher: Breljak und Mühlhoff 2019. Zum Konzept der Subjektivität nach Foucault exemplarisch: Foucault 2007 [1982], 2007a [1984].

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Komplex der medialen Subjektivierung ist ein ethisch und politisch relevanter Schauplatz, denn hier wird verhandelt, dass wir KI-Systeme *freiwillig* und *mit guten Gründen* benutzen und genau dadurch – etwas drastisch formuliert – zu kognitiven Untereinheiten ihrer distribuierten, hybriden Gehirn- und siliziumbasierten Rechennetze werden.

Nur eine Ethik der KI, die auf einem umfassenden Verständnis der Macht von KI-Systemen aufbaut, wird die strukturellen und systemischen Auswirkungen dieser Technologie erfassen können. Eine zeitgemäße Ethik der KI muss deshalb als philosophisches Querschnittprojekt betrieben werden: Sozialphilosophische Theorien und Begriffe wie Macht, Ungleichheit, Diskriminierung, Fairness, Ausbeutung, Subjektivierung, Entmündigung spielen für den Aufriss der ethischen Probleme genauso eine zentrale Rolle wie medienphilosophische Studien zur Funktionsweise vernetzter digitaler Dienste und ihrer Interfaces.[54] Eine solche Ethik der KI steht überdies in fließendem Übergang mit dem Projekt einer kritischen Theorie der KI und des »digitalen Kapitalismus«.[55] Kritische Theorie verweist hier auf einen Begriff und eine philosophische Methodologie von Kritik und vielleicht auch von Aufklärung im Zentrum eines Projekts der Ethik der KI, wenn mit Kritik ein bestimmtes *ethos*, eine philosophisch-ethische Haltung des Fragens und Wirkens gemeint ist. Zentral für diese Haltung ist in der Tradition nach Marx und Foucault das riskante, sich selbst aufs Spiel setzende Fragen nach der Gegenwart, in

54 Siehe exemplarisch: Hadler und Haupt 2016; Kaerlein 2018; Kay 2001 [1989].

55 Vgl. Staab 2019; Crawford 2021; Dyer-Witheford u. a. 2019.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

der wir leben, insofern diese Gegenwart einerseits unser Denken, Fühlen und Handeln prägt und zugleich selbst erst durch unser Denken, Fühlen und Handeln ins Werk gesetzt ist.[56]

Im Fall der KI-Technologie könnte diese Haltung der Kritik dort beginnen, wo auch wir heute Abend begonnen haben: mit der Frage nach der Gegenwart von KI-Technologie. Sie haben gesehen, dass sich das in Feuilletons und Science Fiction produzierte Bild von künstlicher Intelligenz auf Vorstellungen von verkörperten Agenten und Systemen verengt, die tendenziell Zukunftsvisionen sind und mit der aktuellen Realität von KI wenig zu tun haben. Bereits diese Vorstellung von KI-Technologie ist deshalb nicht ethisch neutral. Sie lenkt von der gegenwärtigen Realität dieser Technologie und von den zahlreichen damit verbundenen ethischen Fragen ab. Ich habe versucht, den Standardbeispielen für KI – selbstfahrenden Autos, Schachcomputern, Pflegerobotern – eine Fülle anderer Beispiele entgegenzustellen, die uns nicht als Roboter begegnen, die jedoch schon jetzt gravierende Auswirkungen auf viele von uns und die Gesellschaft im Ganzen haben. Wenn Sie heute Abend eine Botschaft mitnehmen, dann diese: Die Auswirkungen von Big Data und KI betreffen uns schon heute, und zwar jede:n von uns. Sie sind keine Fragen der Zukunft.

56 Vgl. grundsätzlich: Foucault 2007b [1984], 1992 [1978]; Saar 2007; Celikates 2012.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Bibliographie

Ahn, Luis von. 2005. »Human Computation«. Dissertation, School of Computer Science, Carnegie Mellon University.

–. 2006a. »Games with a Purpose«. *Computer* 39 (6): 92–94. https://doi.org/10.1109/MC.2006.196.

–, Reg. 2006b. *Human Computation: Google TechTalk 26 Juy 2006*. https://www.youtube.com/watch?v=tx082gDwGcM.

Ahn, Luis von, und Laura Dabbish. 2004. »Labeling Images with a Computer Game«. In *Proceedings of the 2004 Conference on Human Factors in Computing Systems*, 319–26. Vienna, Austria: ACM Press. https://doi.org/10.1145/985692.985733.

Angwin, Julia, Jeff Larson, Surya Mattu, und Lauren Kirchner. 2016. »Machine Bias«. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Ben-Shahar, Omri. 2019. »Data Pollution«. *Journal of Legal Analysis* 11: 104–59. https://doi.org/10.1093/jla/laz005.

Borodovsky, Jacob T, Lisa A Marsch, und Alan J Budney. 2018. »Studying Cannabis Use Behaviors With Facebook and Web Surveys: Methods and Insights«. *JMIR Public Health and Surveillance* 4 (2). https://doi.org/10.2196/publichealth.9408.

Breljak, Anja, und Rainer Mühlhoff. 2019. »Was ist Sozialtheorie der Digitalen Gesellschaft? – Einleitung«. In *Affekt Macht Netz: Auf dem Weg zu einer Sozialtheorie der digitalen Gesellschaft*, herausgegeben von Rainer Mühlhoff, Anja Breljak, und Jan Slaby, 7–34. Bielefeld: Transcript. https://doi.org/10.14361/9783837644395-001.

Buolamwini, Joy, und Timnit Gebru. 2018. »Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification«. In *Conference on Fairness, Accountability and Trans-*

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

parency, 77–91. PMLR. http://proceedings.mlr.press/v81/buolamwini18a.html.

Cave, Stephen, Kanta Sarasvati Monique Dihal, und Sarah Dillon, Hrsg. 2020. *AI Narratives: A History of Imaginative Thinking about Intelligent Machines*. Oxford: Oxford University Press.

Celikates, Robin. 2012. »Karl Marx: Critique as Emancipatory Practice«. In *Conceptions of Critique in Modern and Contemporary Philosophy*, herausgegeben von Karin de Boer und Ruth Sonderegger, 101–18. London: Palgrave Macmillan UK. https://doi.org/10.1057/9780230357006.

Coeckelbergh, Mark. 2020. *AI Ethics*. Cambridge, MA: The MIT Press.

Confessore, Nicholas, und Danny Hakim. 2017. »Data Firm Says ›Secret Sauce‹ Aided Trump; Many Scoff«. *The New York Times*, 6. März 2017. https://www.nytimes.com/2017/03/06/us/politics/cambridge-analytica.html.

Crawford, Kate. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.

Davies, Harry. 2015. »Ted Cruz Using Firm That Harvested Data on Millions of Unwitting Facebook Users«. *The Guardian*, 11. Dezember 2015. https://www.theguardian.com/us-news/2015/dec/11/senator-ted-cruz-president-campaign-facebook-user-data.

Denecke, K., P. Bamidis, C. Bond, E. Gabarron, M. Househ, A. Y. S. Lau, M. A. Mayer, M. Merolli, und M. Hansen. 2015. »Ethical Issues of Social Media Usage in Healthcare«. *Yearbook of Medical Informatics* 10 (1): 137–47. https://doi.org/10.15265/IY-2015-001.

Dyer-Witheford, Nick, Atle Mikkola Kjøsen, und James Steinhoff. 2019. *Inhuman Power: Artificial Intelligence and the Future of Capitalism*. Pluto Press.

Eubanks, Virginia. 2017. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. First Edition. New York, NY: St. Martin's Press.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Facebook. 2021. »An Update On Our Use of Face Recognition«. *Meta Company Blog* (blog). 2. November 2021. https://about.fb.com/news/2021/11/update-on-use-of-face-recognition/.

Foucault, Michel. 1983 [1976]. *Der Wille zum Wissen: Sexualität und Wahrheit 1*. Übersetzt von Ulrich Raulff und Walter Seitter. Frankfurt am Main: Suhrkamp.

–. 1992 [1978]. *Was ist Kritik?* Übersetzt von Walter Seitter. Internationaler Merve-Diskurs 167. Berlin: Merve-Verl.

–. 2007a [1984]. »Foucault (Maurice Florence)«. In *Ästhetik der Existenz: Schriften zur Lebenskunst*, herausgegeben von Daniel Defert und François Ewald, 220–25. Frankfurt am Main: Suhrkamp.

–. 2007 [1982]. »Subjekt und Macht«. In *Ästhetik der Existenz: Schriften zur Lebenskunst*, herausgegeben von Daniel Defert und François Ewald, 81–104. Frankfurt am Main: Suhrkamp.

–. 2007b [1984]. »Was ist Aufklärung?« In *Ästhetik der Existenz: Schriften zur Lebenskunst*, herausgegeben von Daniel Defert und François Ewald, 171–90. Frankfurt am Main: Suhrkamp.

Grassegger, Hannes, und Mikael Krogerus. 2016. »Ich habe nur gezeigt, dass es die Bombe gibt«. *Das Magazin* 48.

Guthrie, Katherine A., Bette Caan, Susan Diem, Kristine E. Ensrud, Sharon R. Greaves, Joseph C. Larson, Katherine M. Newton, Susan D. Reed, und Andrea Z. LaCroix. 2019. »Facebook Advertising for Recruitment of Midlife Women with Bothersome Vaginal Symptoms: A Pilot Study«. *Clinical Trials* 16 (5): 476–80. https://doi.org/10.1177/1740774519846862.

Hadler, Florian, und Joachim Haupt. 2016. *Interface Critique*. https://doi.org/10.13140/RG.2.2.27453.05604.

Haugeland, John. 1985. *Artificial Intelligence: The Very Idea*. Cambridge, Mass: MIT Press.

Hern, Alex. 2018. »Cambridge Analytica: How Did It Turn Clicks into Votes?« *The Guardian*, 6. Mai 2018. https://www.the

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

guardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie.

Hutchinson, Andrew. 2022. »Meta Reiterates the Value of Personalized Ad Tracking in New Ad Campaign«. *Social Media Today*, 5. April 2022. https://www.socialmediatoday.com/news/meta-reiterates-the-value-of-personalized-ad-tracking-in-new-ad-campaign/621642/.

Joque, Justin. 2022. *Revolutionary Mathematics: Artificial Intelligence, Statistics and the Logic of Capitalism*. London New York: Verso.

Kaerlein, Timo. 2018. *Smartphones als digitale Nahkörpertechnologien: Zur Kybernetisierung des Alltags*. Bielefeld: Transcipt Verlag. https://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=5504252.

Kay, Alan. 2001 [1989]. »User Interface: A Personal View«. In *multiMEDIA – From Wagner to Virtual Reality*, herausgegeben von Randall Packer und Ken Jordan, 121–31. New York: W.W.Norton.

Keddell, Emily. 2015. »The Ethics of Predictive Risk Modelling in the Aotearoa/New Zealand Child Welfare Context: Child Abuse Prevention or Neo-Liberal Tool?« *Critical Social Policy* 35 (1): 69–88. https://doi.org/10.1177/0261018314543224.

Kosinski, Michal, David Stillwell, und Thore Graepel. 2013. »Private Traits and Attributes Are Predictable from Digital Records of Human Behavior«. *Proceedings of the National Academy of Sciences* 110 (15): 5802–5. https://doi.org/10.1073/pnas.1218772110.

Lippert, John. 2014. »ZestFinance Issues Small, High-Rate Loans, Uses Big Data to Weed Out Deadbeats«. *Washington Post*, 11. Oktober 2014. https://www.washingtonpost.com/business/zestfinance-issues-small-high-rate-loans-uses-big-data-to-weed-out-deadbeats/2014/10/10/e34986b6-4d71-11e4-aa5e-7153e466a02d_story.html.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Martini, Mario, Jonas Botta, David Nink, Michael Kolain, und Bertelsmann Stiftung. 2020. »Automatisch erlaubt?: Fünf Anwendungsfälle algorithmischer Systeme auf dem juristischen Prüfstand«. *Impuls Algorithmenethik*. https://doi.org/10.11586/2019067.

Matthews, Gerald, Ian J. Deary, und Martha C. Whiteman. 2003. *Personality Traits*. 2nd ed. Cambridge, U.K. ; New York: Cambridge University Press.

Matzner, Tobias. 2019. »Autonomy Trolleys und andere Probleme: Konfigurationen künstlicher Intelligenz in ethischen Debatten über selbstfahrende Kraftfahrzeuge«. *Zeitschrift für Medienwissenschaft* 21 (2): 46–55.

MD Connect. 2017. »Social Media & Clinical Trial Recruitment [Whitepaper]«. https://cdn2.hubspot.net/hubfs/291282/documents/Gated_Content/White%20Paper%20-%20Clinical%20Trial%20-%20Social%20Media%20Patient%20Recruitment.pdf.

Merchant, Raina M., David A. Asch, Patrick Crutchley, Lyle H. Ungar, Sharath C. Guntuku, Johannes C. Eichstaedt, Shawndra Hill, Kevin Padrez, Robert J. Smith, und H. Andrew Schwartz. 2019. »Evaluating the Predictability of Medical Conditions from Social Media Posts«. *PLOS ONE* 14 (6): e0215476. https://doi.org/10.1371/journal.pone.0215476.

Misselhorn, Catrin. 2018. *Grundfragen der Maschinenethik*. 4. Aufl. Stuttgart: Reclam.

Mittelstadt, Brent, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, und Luciano Floridi. 2016. »The Ethics of Algorithms: Mapping the Debate«. *Big Data and Society* 3 (2). https://doi.org/10.1177/2053951716679679.

Mühlhoff, Rainer. 2019a. »Big Data is Watching You. Digitale Entmündigung am Beispiel von Facebook und Google«. In *Affekt Macht Netz: Auf dem Weg zu einer Sozialtheorie der digitalen Gesellschaft*, herausgegeben von Rainer Mühlhoff, Anja Breljak,

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

und Jan Slaby, 81–107. Bielefeld: Transcript. https://doi.org/10.14361/9783837644395-004.

–. 2019b. »Menschengestützte Künstliche Intelligenz: Über die soziotechnischen Voraussetzungen von Deep Learning«. *Zeitschrift für Medienwissenschaft* 21 (2): 56–64. https://doi.org/10.25969/mediarep/12633.

–. 2020a. »Human-Aided Artificial Intelligence: Or, How to Run Large Computations in Human Brains? Toward a Media Sociology of Machine Learning«. *New Media & Society* 22 (10): 1868–84. https://doi.org/10.1177/1461444819885334.

–. 2020b. »Automatisierte Ungleichheit: Ethik der Künstlichen Intelligenz in der biopolitischen Wende des Digitalen Kapitalismus«. *Deutsche Zeitschrift für Philosophie* 68 (6): 867–90. https://doi.org/10.1515/dzph-2020-0059.

–. 2021. »Predictive Privacy: Towards an Applied Ethics of Data Analytics«. *Ethics and Information Technology*. https://doi.org/10.1007/s10676-021-09606-x.

–. 2022. »Prädiktive Privatheit: Kollektiver Datenschutz im Kontext von Big Data und KI«. In *Künstliche Intelligenz, Demokratie und Privatheit*, herausgegeben von Michael Friedewald, Alexander Roßnagel, Jessica Heesen, Nicole Krämer, und Jörn Lamla, 31–58. Nomos Verlagsgesellschaft mbH & Co. KG. https://doi.org/10.5771/9783748913344-31.

Mühlhoff, Rainer, und Hannah Ruschemeier. 2022. »Predictive Analytics und DSGVO: Ethische und rechtliche Implikationen«. In *Telemedicus – Recht der Informationsgesellschaft: Tagungsband zur Sommerkonferenz 2022*, herausgegeben von Hans-Christian Gräfe und Telemedicus e.V., 38–67. Frankfurt am Main: Deutscher Fachverlag.

Mühlhoff, Rainer, und Theresa Willem. 2022. »Social Media Advertising for Clinical Studies: Ethical and Data Protection Implications of Online Targeting«. *Big Data & Society* (im Er-

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

scheinen). Pre-Print: https://rainermuehlhoff.de/media/publications/muehlhoff-willem-2022-SMACS-preprint.pdf.

Ng, Andrew Y, Reg. 2017. *Andrew Ng: Artificial Intelligence Is the New Electricity: Talk at Stanford Graduate School of Business*. https://www.youtube.com/watch?v=21EiKfQYZXc.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

Nosthoff, Anna-Verena, und Felix Maschewski. 2019. »We have to Coordinate the Flow‹ oder: Die Sozialphysik des Anstoßes. Zum Steuerungs- und Regelungsdenken neokybernetischer Politiken«. In *Steuern und Regeln*, herausgegeben von Alexander Friedrich, Petra Gehring, Christoph Hubig, Andreas Kaminski, und Alfred Nordmann, 39–54. Nomos Verlagsgesellschaft mbH & Co. KG. https://doi.org/10.5771/9783845296548-39.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. First edition. New York: Crown.

Rosenberg, Matthew, Nicholas Confessore, und Carole Cadwalladr. 2018. »How Trump Consultants Exploited the Facebook Data of Millions«. *The New York Times*, 17. März 2018. https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html.

Ruschemeier, Hannah. 2021. »Kollektiver Rechtsschutz und strategische Prozessführung gegen Digitalkonzerne«. *MMR* 24 (12): 942–46.

Saar, Martin. 2007. *Genealogie als Kritik: Geschichte und Theorie des Subjekts nach Nietzsche und Foucault*. Frankfurt am Main ; New York: Campus Verlag.

Sellers, Frances Stead. 2015. »Cruz Campaign Paid $750,000 to ›Psychographic Profiling‹ Company«. *Washington Post*, 19. Oktober 2015. https://www.washingtonpost.com/politics/cruz-cam

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

paign-paid-750000-to-psychographic-profiling-company/2015/10/19/6c83e508-743f-11e5-9cbb-790369643cf9_story.html.

Staab, Philipp. 2019. *Digitaler Kapitalismus: Markt und Herrschaft in der Ökonomie der Unknappheit*. Suhrkamp Verlag.

Tufekci, Zeynep. 2014. »Engineering the Public: Big Data, Surveillance and Computational Politics«. *First Monday*. https://doi.org/10.5210/fm.v19i7.4901.

Wagner, Gerhard, und Horst Eidenmuller. 2019. »Down by Algorithms: Siphoning Rents, Exploiting Biases, and Shaping Preferences: Regulating the Dark Side of Personalized Transactions«. *U. Chi. L. Rev.* 86: 581.

Weiser, Mark. 1991. »The Computer for the 21st Century«. *ACM SIGMOBILE Mobile Computing and Communications Review* 3: 3–11.

Wisk, Lauren E., Eliza B. Nelson, Kara M. Magane, und Elissa R. Weitzman. 2019. »Clinical Trial Recruitment and Retention of College Students with Type 1 Diabetes via Social Media: An Implementation Case Study«. *Journal of Diabetes Science and Technology* 13 (3): 445–56. https://doi.org/10.1177/1932296819839503.

Young, Iris Marion. 1994. »Gender as Seriality: Thinking about Women as a Social Collective«. *Signs* 19 (3): 713–38. https://www.jstor.org/stable/3174775.

Zuboff, Shoshana. 2015. »Big Other: Surveillance Capitalism and the Prospects of an Information Civilization«. *Journal of Information Technology* 30 (1): 75–89. https://doi.org/10.1057/jit.2015.5.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über https://dnb.de abrufbar.

ISSN 2198-6258
ISBN 978-3-8471-1552-6

Veröffentlichungen des Universitätsverlags Osnabrück erscheinen bei V&R unipress.

Druck und Bindung: Memminger MedienCentrum, Fraunhoferstr. 19, 87700 D-Memmingen
Printed in the EU.

ISBN Print: 9783847115526 – ISBN E-Lib: 9783737015523