Rainer Mühlhoff*

Automatisierte Ungleichheit

Ethik der Künstlichen Intelligenz in der biopolitischen Wende des Digitalen Kapitalismus

https://doi.org/10.1515/dzph-2020-0059

Abstract: This paper sets out the notion of a current "biopolitical turn of digital capitalism" resulting from the increasing deployment of AI and data analytics technologies in the public sector. With applications of AI-based automated decisions currently shifting from the domain of business to customer (B2C) relations to government to citizen (G2C) relations, a new form of governance arises that operates through "algorithmic social selection". Moreover, the paper describes how the ethics of AI is at an impasse concerning these larger societal and socio-economic trends and calls for an ethics of AI that includes, and acts in close alliance with, social and political philosophy. As an example, the problem of Predictive Analytics is debated to make the point that data-driven AI (Machine Learning) is currently one of the main ethical challenges in the ethics of AI.

Keywords: ethics of AI, Machine Learning, Predictive Analytics, biopolitics, social selection, digital capitalism, automated decision-making, critical theory, social philosophy

1 Einleitung

In der Ethik der Künstlichen Intelligenz gibt es seit längerem zwei unterschiedliche Herangehensweisen. Die eine, auch unter dem Begriff der Maschinen-Ethik bekannt, stellt die Frage nach dem ethischen Agieren künstlich-intelligenter, autonomer Systeme.[1] Ethik betrifft hier das Handeln der Systeme, insofern sie in die „menschliche" Gesellschaft eingebettet sind und *als darin Handelnde* möglicherweise Auswirkungen auf andere Menschen, Lebewesen oder andere künstli-

1 Vgl. Anderson/Anderson (2011) u. Misselhorn (2018).

***Kontakt: Rainer Mühlhoff,** Technische Universität Berlin, Cluster Science of Intelligence, Marchstraße 23, 10587 Berlin; muehlhoff@tu-berlin.de

che autonome Systeme haben. Das ethische Subjekt dieser Debatte sind die autonomen Maschinen selbst und mittelbar ihre ErschafferInnen, die dafür sorgen müssen, dass die Maschinen ethisch agieren. Auch die weithin mit KI-Ethik assoziierten Dilemmata vom Typ des Trolley-Problems fallen in die Maschinen-Ethik: Steht die Maschine vor einer ethischen Abwägungssituation? Nach welchen Prinzipien wird sie entscheiden?[2] Wie stellen wir sicher, dass sie sich dabei an den gesellschaftlichen Werten orientiert (und welche genau sind das)?[3]

Eine zweite Herangehensweise ist die Technikethik mit ihren verschiedenen Bereichsethiken, zu denen auch eine Technikethik der künstlichen Intelligenz zählt.[4] Sie steht methodisch und systematisch in Zusammenhang mit Technikfolgenabschätzung, Science and Technology Studies und Verantwortlicher Technikgestaltung.[5] In vielerlei Hinsicht der Medizinethik ähnlich[6], sind die ethischen Subjekte dieser Stoßrichtung nicht die technologischen Artefakte, sondern jene Menschen und gesellschaftlichen Instanzen, die an der Erschaffung, Beforschung, Finanzierung, Umsetzung oder Inbetriebnahme der Technologien beteiligt sind. Die Frage der Technikethik betrifft „Responsible Research and Innovation" (RRI) auf Basis einer ethischen Bewertung der möglichen gesellschaftlichen Veränderungen durch neue Technologie.[7] Ein ethischer Diskurs dieser Form wendet sich an die Gesellschaft und ihre EntscheiderInnen, PolitikerInnen, ForscherInnen und ökonomischen Akteure, die technologische Neuerungen auf den Weg bringen, anwenden und potenziell regulieren können. Dabei gilt es erstens, überhaupt ein Bewusstsein für die möglichen Konsequenzen technologischer Errungenschaften zu erarbeiten, und zweitens, in den hypothetischen Raum denkbarer Auswirkungen durch Regulierungen und Wertsetzungen einzugreifen.

Dieser Artikel zielt auf eine ethische Thematisierung jener KI-Technologien, die aktuell die deutlichsten Auswirkungen auf Gesellschaft, Politik und Wirtschaft haben: datenbasierte Prognosesysteme in der Gestalt von prädiktiver Analytik, algorithmischer Profilbildung, verhaltensbasiertem Targeting, datenbasierten Klassifikations- und Scoring-Verfahren, Nachrichtenfilterung (news recommender systems) etc. Hierbei handelt es sich meist um Anwendungen maschineller Lernverfahren, die in großen Datenmengen (Big Data) Korrelationen analysieren und „Muster erkennen" können, um dadurch zukünftiges Verhalten oder Ereig-

2 Vgl. zur kritischen Diskussion Etzioni/Etzioni (2017) u. Matzner (2019).
3 Zum sog. „Value Alignment Problem" vgl. Wong/Simon (2020).
4 Vgl. Dignum (2019).
5 Vgl. Grunwald (2013).
6 Vgl. Véliz (2019).
7 Vgl. Owen et al. (2013).

nisse zu prognostizieren, so dass das Verhalten Einzelner vorbeugend antizipiert oder sogar beeinflusst werden kann.[8] Es wird zunehmend darauf hingewiesen, dass der Einsatz dieser Technologien zu einer „Automatisierung von Ungleichheit" und neuen Formen der sozialen Selektion führt, die mit den Wertschöpfungsformen des digitalen Kapitalismus verwoben sind.[9] Sie gestatten es, in zahlreichen Lebens- und Anwendungsbereichen Individuen anhand von Verhaltensähnlichkeiten in prognostische Gruppen einzuteilen und auf dieser Grundlage unterschiedlich zu behandeln.

Um diesen Gegenstandsbereich zu adressieren, werde ich mich grob der zweiten der oben genannten Herangehensweisen von Ethik der KI anschließen. Zugleich ist es das Ziel dieses Beitrags, programmatisch herauszuarbeiten, um welche Perspektiven und Querverbindungen eine Ethik der KI erweitert werden sollte, um der komplexen gesellschaftlichen Verwobenheit dieser Technologien, insbesondere mit aktuellen Wirtschaftsformen, gerecht zu werden. Nur so wird Ethik der KI den gesellschaftlichen Einfluss entfalten, der ihr angesichts der Tragweite dieses Gegenstandsbereichs gebührt. Das ist nötig, weil Ethik der KI, trotz der hohen Aufmerksamkeit, die sie zur Zeit genießt, zunehmend in eine Sackgasse gelangt: Im Jahr 2020 ist es nicht mehr opportun, künstliche Intelligenz zu betreiben, ohne sich dabei öffentlichkeitswirksam auf eine „ethische" Vorgehensweise zu verpflichten. Nahezu 100 Richtlinienpapiere, Kriterienlisten, Whitepapers und Absichtserklärungen wurden in den letzten vier Jahren von Firmen, Regierungen, NGOs und Forschungseinrichtungen zu diesem Thema herausgebracht.[10] KI-EthikerInnen zählen nach Auskunft von Unternehmensberatungen zu den wichtigsten Fachkräften für Unternehmen, die im Bereich KI erfolgreich werden möchten.[11] Doch es mehren sich die Erkenntnisse, dass KI-Ethik zwar populär, aber keineswegs auch wirkungsvoll ist; der Begriff Ethik wird vielmehr von wirtschaftlichen Akteuren und PR-Strategen angeeignet, die sich damit in ein günstiges Licht rücken können. Daraus wird immer öfter geschlussfolgert, dass man keine Ethik, sondern Gesetze, Regulierungen und Verpflichtungen auf Menschenrechte benötige, um den Gefahren durch künstliche Intelligenz nicht bloß weiche Selbstverpflichtungen entgegenzustellen.[12]

Dieses Problem ist in der Tat gravierend, doch die Lösung besteht nicht in der Abwendung von Ethik, sondern in der Schärfung ihres Verhältnisses zu –

8 Vgl. Mittelstadt et al. (2016) u. O'Neil (2016).
9 Eubanks (2017) u. O'Neil (2016).
10 Vgl. Jobin et al. (2019).
11 Vgl. Wong/Simon (2020).
12 Vgl. Mittelstadt (2019), Saslow/Lorenz (2019) u. Wagner (2018).

ebenso notwendiger – politischer Regulierung und Selbstverpflichtung beteiligter Akteure. Denn nur ein ethischer Diskurs kann das grundlegende Problem, zu dem wir uns gesellschaftlich verhalten müssen, in seiner Tiefe befragen – jenseits der Pragmatik von Kriterienkatalogen und Checklisten. Im Fall prädiktiver Analytik besteht die grundsätzliche Frage darin, ob wir Menschen als individuelle Risikofaktoren behandeln möchten, die es zu erkennen und zu verwalten gilt, anstatt als prinzipiell gleiche Mitglieder einer Solidargemeinschaft, in der Risiken verteilt und gemeinsam geschultert werden. Ich werde in diesem Beitrag eine ethische Herangehensweise an dieses Problem skizzieren, die an die Technikethik anschließt und zugleich über sie hinausgeht. Dazu gehe ich in drei Schritten vor:

1. Klärung des Gegenstandsbereichs: Der massenhafte Einsatz des maschinellen Lernens im Kontext von Big Data, prädiktiver Analytik und Verhaltensprognostik ist ein Feld, in dem KI-Technologie aktuell bereits flächendeckende gesellschaftliche Auswirkungen zeigt. Anders als autonome Roboter oder selbstfahrende Autos, die weitaus weniger realisiert und verbreitet sind, ist das gesellschaftliche Bewusstsein von diesen Theorien noch nicht so ausgeprägt, was auch an ihrer Unsichtbarkeit liegt. In Abschnitt 2 werde ich deshalb zunächst die hier betrachtete Technologie und ihre Anwendungen genau beschreiben und für ihre philosophische Besprechung relevante Begriffe einführen.
2. Allianz mit der Sozialphilosophie: Um prädiktive Analytik (PA) fundiert zu diskutieren, muss Ethik umfassend sozialtheoretisch und polit-ökonomisch informiert sein. Wie ich in Abschnitt 3 argumentiere, bedeutet das, in die ethische Betrachtung eine Analyse der Machtmuster, Subjektivierungsformen, Kapitalakkumulationsweisen und lokalen wie globalen Ungleichheiten einzubeziehen, die mit KI-Technologie in Zusammenhang stehen. Die in der Algorithmen-Ethik geläufige Thematisierung von Biases, Opazität und Diskriminierung[13] geht zwar in die richtige Richtung, könnte jedoch an Wirksamkeit hinzugewinnen, wenn sie auch zum materiellen, global-ökonomischen Ursprung dieser Phänomene vordränge.
3. Bestimmung des ethischen Subjekts: Eine wirkungsvolle Ethik der PA im speziellen und der KI im allgemeinen bekennt sich zur Rolle ethischer Diskurse für eine kollektive Praxis der Bildung, Reflexion und Kritik. Damit ist die selbstreflexive und mitgestaltende Arbeit einer politisch-demokratischen Gemeinschaft gemeint. Ethik der KI muss deshalb die beteiligten und betroffenen Subjekte *ansprechen*, sie als ethische Agenten aktivieren und in die Pflicht nehmen. Wie ich in Abschnitt 4 argumentiere, gehören dazu neben

13 Vgl. Mittelstadt et al. (2016); Mittelstadt (2019); sowie Friedman/Nissenbaum (1996).

EntscheiderInnen und DesignerInnen neuer Technologien vor allem auch die NutzerInnen und somit quasi *wir alle* – eine Gruppe moralischer Agenten, die in der ethischen Thematisierung von KI bisher weitestgehend vernachlässigt wird.

2 Gegenstandsbereich: Prädiktive Analytik und algorithmisches Entscheiden

Künstliche Intelligenz ist seit einigen Jahren wieder ein „gehyptes Thema". Nachrichten über reale technische Erfolge mischen sich in der öffentlichen Debatte mit vagen Zukunftsvisionen, Geschäftsideen, Utopien, fiktiven Beiträgen und Science Fiction. Ethik der KI steht deshalb vor dem Problem, eine imaginäre Wirkungsdimension (gesellschaftliche Auswirkungen von Visionen und Fiktionen) von realen Effekten (Auswirkungen von jetzt oder demnächst eingesetzten Technologien) abzugrenzen. Einige Debatten der Maschinenethik, die sich auf das ethische Agieren autonomer Systeme beziehen, bewegen sich hinsichtlich ihres Gegenstandsbereichs deutlich im Bereich des Visionären.[14] Auch wenn sich solche Beiträge oft als neutrale Gedankenübungen verstehen mögen, besitzen sie eine problematische Diskursfunktion: Sie stützen und stabilisieren den Hype sowie das kulturelle Imaginäre von autonom agierenden KI-Systemen und lenken die Aufmerksamkeit nicht selten auf Randfälle und Spezialprobleme, die an der täglichen Realität, die durch KI-Anwendungen längst erzeugt wird, vorbeigehen.[15]

Weniger im Blick der öffentlichen Diskussion stehen jene KI-Technologien, die in den Gegenstandsbereich der Algorithmen-Ethik fallen.[16] Damit sind KI-basierte Prognosesysteme, algorithmisches Entscheiden, Profiling und prädiktive Analytik gemeint. Ich werde im Folgenden zunächst definieren, was ich unter KI-basierten Prognosesystemen verstehe und anschließend argumentieren, warum diese KI-Anwendungen aufgrund ihres aktuell hohen Verbreitungsgrades im Fokus dieses Beitrags stehen.

14 Vgl. Bostrom/Yudkowsky (2014).
15 Vgl. Etzioni/Etzioni (2017) sowie Matzner (2019).
16 Mittelstadt et al. (2016); Tufekci (2015); Zwitter (2014).

2.1 KI-basierte Prognosesysteme und prädiktive Analytik

Unter der Bezeichnung „KI-basierte Prognosesysteme“ lassen sich verschiedene Anwendungen von Machine Learning-(ML-)Modellen für die Vorhersage von Verhalten, Risiken, Interessen oder Eigenschaften von Individuen zusammenfassen. Hierbei bilden ML-Algorithmen einen Spezialfall statistischer Auswertungen, deshalb fallen KI-basierte Prognosesysteme in den allgemeineren Bereich der statistischen Prognosesysteme. Ein bekanntes Beispiel liefert der Empfehlungsdienst des Online-Händlers Amazon, der den BenutzerInnen automatisiert anhand bisher gekaufter oder betrachteter Artikel weitere potenziell interessante Artikel vorschlägt.

Wenn weniger das informationstheoretische Gesamtsystem benannt werden soll, sondern seine Funktion der Prognosestellung, dann spricht man auch von einer „prädiktiven Analyse“. Hierbei handelt es sich um eine mathematische Funktion oder einen Algorithmus

$$P_W : D_i \to A_i,$$

der von einem Bestand W an Erfahrungswissen (Trainingsdaten) abhängt und darauf aufbauend für ein Input-Datum D_i eine Vorhersage A_i zurückgibt.[17] Hierbei sind D_i typischerweise die Informationen, die über ein Individuum oder einen Fall i vorliegen; A_i enthält dann eine Prognose bestimmter Eigenschaften, die über i unbekannt sind. Im Fall der Produktempfehlungen auf Amazon bestünde das Erfahrungswissen W zum Beispiel aus der gesamten Kaufhistorie aller NutzerInnen der Plattform, kombiniert mit allen personenbezogenen und Nutzungsdaten, die über die KundInnen vorliegen. In diesem Datensatz lassen sich, zum Beispiel mittels maschinellem Lernen, Muster und Korrelationen im Kaufverhalten von Million KundInnen analysieren. Das Resultat wäre ein Modell P_W, das anhand der über eine konkrete NutzerIn i vorliegenden Daten D_i (ihre individuelle Kaufhistorie, potenziell weitere personenbezogene und Nutzungsdaten) eine Liste A_i von Produkten ermittelt, die sie wahrscheinlich in der nahen Zukunft kaufen würde.

Entscheidend für eine prädiktive Analyse ist, dass P_W von dem Erfahrungswissen W über viele empirische Fälle abhängt: Prädiktive Analytik verfährt nach einer induktiven Schlussweise, indem sie statistische Inferenzen, die über eine

17 Vgl. zu den mathematischen Grundlagen Grindrod (2014).

große Grundgesamtheit ermittelt werden, auf den Einzelfall überträgt.[18] Prädiktiver Analytik liegt deshalb ein radikal empirisches Prinzip zugrunde, das den vorgelegten Einzelfall letztlich anhand einer Vergleichsoperation mit den in der Wissensbasis *W* enthaltenen Fällen einschätzt und nicht etwa deduktiv das Resultat A_i allein aus den Einzelfalleigenschaften D_i ableitet.[19]

Produktempfehlung ist ein vergleichsweise harmloser Einsatzbereich solcher Prognosesysteme. Ein weiteres Beispiel, das in den verwandten Bereich des „targeted advertising" fällt, ist die Erkennung schwangerer Supermarkt-KundInnen anhand ihres Kaufverhaltens, um sie mit abgestimmter Werbung (z. B. für Baby-Produkte) versorgen zu können.[20] Hieran lässt sich gut studieren, wie solche Systeme trainiert werden: Kann man – etwa über Kundenkarten oder Kreditkartendaten – das Kaufverhalten von KundInnen über lange Zeit verfolgen, dann sind werdende Eltern *retrospektiv* mit hoher Treffsicherheit erkennbar, sobald sie (mutmaßlich ab Geburt des Kindes) einschlägige Baby-Produkte kaufen. Nach diesem Kriterium können Schwangere ex post in den Daten identifiziert werden, um ihr Kaufverhalten in der Zeit vor Eintritt der Elternschaft mit dem von Nicht-Schwangeren zu vergleichen und auf aussagekräftige „Muster" oder Prädiktoren hin zu untersuchen. Ist ein solches Modell einmal erstellt worden, lässt es sich dann dafür verwenden, werdende Eltern noch während der Schwangerschaft zu erkennen und vorausblickend mit spezifischer Werbung anzusprechen.

Ein ähnliches Prinzip liegt den Anwendungen KI-basierter Prognostik im Credit Scoring und bei der individuellen Bepreisung von Versicherungen zugrunde.[21] Auch hier können Informationen über Zahlungsausfälle oder Schadensfälle des bestehenden Kundenstamms genutzt werden, um retrospektiv Prädiktoren für diese Ereignisse zu suchen, die bei neuen KundInnen zur Anwendung gebracht werden. Sogenannte „Payday-Lending"-Anbieter wie die amerikanische Firma ZestFinance oder die deutsche Firma Kreditech haben sich auf die Bereitstellung von Krediten und Finanzprodukten „[for] the world's unbanked" spezialisiert.[22] Sie verwenden Risikomodelle auf Grundlage prädiktiver Analytik, um „bankenlose" KundInnen, die auf dem klassischen Finanzmarkt als nicht kreditfähig gelten, mit Krediten zu versorgen. Oft in Situationen der Ausweglosigkeit

18 Der Bezug auf eine große Datenmenge vieler Fälle steckt in dem Wort „Analytik": Teil des Verfahrens besteht in der daten-analytischen Ermittlung einer Gesetzmäßigkeit.

19 Dies ist ein entscheidender konzeptueller Punkt, der in der Debatte um „Algorithmen-Ethik", „algorithmisches Entscheiden" etc. oft unscharf bleibt: Dort ist häufig von „Algorithmen" per se die Rede, was aber unnötig allgemein ist; vgl. Mittelstadt et al. (2019).

20 Duhigg (2012).

21 Vgl. Hurley/Adebayo (2017) sowie O'Neil (2016), 161 ff.

22 O'Dwyer (2018).

und als letztes Mittel konsultiert, nötigen diese Firmen den KreditbewerberInnen bei der Antragstellung, die nicht zufällig ausschließlich online möglich ist, weitreichende Zugriffe auf persönliche Daten ab, darunter ihre Profile in sozialen Netzwerken. Über ZestFinance wurde bekannt, dass auch die Schreibfehler im Online-Antrag ermittelt und die Zeit, die die BewerberIn zum Durchlesen der Vertragsdaten benötigt, gemessen werden. All diese Datenpunkte werden zu dem Input D_i einer prädiktiven Analyse zusammengefasst, die dann ermittelt, ob eine BewerberIn einen Kredit erhält, und wenn ja, zu welchen Zinskonditionen – diese liegen nicht selten bei mehr als 300 % p. a. (sic!).[23]

Der klassische FICO Score in den USA und der Schufa-Score in Deutschland zur Einschätzung des individuellen Kreditrisikos verwendeten ursprünglich nur Informationen aus der eigenen Kreditgeschichte einer Person.[24] Im Unterschied dazu beruht das Credit Scoring mittels prädiktiver Analytik auf einer lateralen Vergleichsoperation mit vielen anderen Individuen. Dies ist eine charakteristische Eigenschaft von prädiktiver Analytik: Sie teilt die Individuen nach einem „People-like-you"-Prinzip[25] in Kategorien ein, indem verschiedene Individuen und Fälle anhand der über sie bekannten Daten (hier: Social Media, Lesegeschwindigkeit, Schreibfehler, ...) miteinander in Verbindung gebracht werden, das heißt, automatisiert verglichen, nach Ähnlichkeit gruppiert und anschließend auch in Bezug auf die (noch) nicht über sie bekannten Parameter (hier: zukünftiges Rückzahlungsverhalten) als ähnlich angenommen werden. Charakteristisch für die komplexen Verfahren des maschinellen Lernens ist, dass die Kriterien, nach denen Ähnlichkeitsgruppierungen vorgenommen werden, nicht von außen vorgegeben werden und nicht auf nachvollziehbare, zum Beispiel demographische Parameter abbildbar sein müssen, sondern von den Algorithmen selbst erlernt werden. Sie lassen sich deshalb *prinzipiell* weder kontrollieren noch erklären, vielmehr sind sie auch für die EntwicklerInnen „opak", „ad hoc" und „ephemer".[26]

Mit einer ähnlichen Vorgehensweise gelingt es, anhand von Facebook-Daten vorherzusagen, ob eine NutzerIn an Krankheiten wie Depression, Psychosen, Diabetes oder Bluthochdruck leidet, wie MedizinerInnen von der University of Pennsylvania gezeigt haben.[27] Solche prädiktiven Analysen sind bei Versicherun-

23 Vgl. Lippert (2014).
24 Vgl. O'Neil (2016), 141 ff., sowie Hurley/Adebayo (2017).
25 Vgl. O'Neil (2016), 145.
26 Mittelstadt (2017).
27 Vgl. Merchant et al. (2019).

gen von großem Interesse, weil sie eine individuelle Risikobemessung erlauben.[28] Facebook selbst hat bekanntgegeben, mittels künstlicher Intelligenz suizidale NutzerInnen anhand ihrer Postings erkennen zu können und in akuten Fällen automatisiert die Behörden zu informieren.[29] In der Kriminologie wird maschinelles Lernen für die Abschätzung von Rückfallwahrscheinlichkeiten (Recidivism Scoring) verwendet.[30] Im Human-Resource-Management ist das KI-basierte Vorsortieren von BewerberInnen bei Jobausschreibungen mittels sogenannter „hiring algorithms" ein großes Thema;[31] in den USA werden solche Systeme bereits heute bei einer Mehrheit der Einstellungsverfahren verwendet.[32] Und schließlich ist auch der Einsatz von prädiktiver Analytik im Bereich der Bildungsarbeit ein bedeutendes Anwendungsfeld: „Educational Data Mining" und „Learning Analytics" untersuchen Lernprofile von SchülerInnen und Studierenden, klassifizieren sie nach Stärken und Schwächen und treffen Vorhersagen über ihre Leistungsfähigkeit.[33] Es ist bekannt, dass solche Analysen auch mit Anwendungen im Human Resource Management verknüpft werden können.[34]

2.2 Strukturalität KI-basierter Prognosesysteme

Prädiktive Analytik und KI-basierte Prognosesysteme finden gemessen an ihrem gesellschaftlichen Einfluss noch zu wenig Aufmerksamkeit in ethischen und kritischen öffentlichen Debatten. Das dürfte auch daran liegen, dass sich diese KI-Technologien nicht als materielles Gegenüber der Menschen präsentieren, das heißt, nicht in Form eines autonom handelnden, körperlich gegenwärtigen Systems auf gleicher Interaktionsebene in Erscheinung treten. Es handelt sich bei diesen KIs vielmehr um *Strukturmerkmale digitaler Räume und Prozesse*. Diese KI-Systeme bilden mehr und mehr den technologischen Rahmen von sozialen, kommunikativen, ökonomischen oder politischen Beziehungen: Sie schaffen die Bedingungen unserer Handlungsmöglichkeiten und unseres Bewusstseins von Handlungsmöglichkeiten überhaupt.

28 Vgl. O'Neil (2016), 161 ff.
29 Vgl. Goggin (2019).
30 Vgl. Hao (2019).
31 Bogen (2019).
32 Vgl. O'Neil (2016), 108 u. 148.
33 Baker/Inventado (2014).
34 Vgl. O'Neil (2016).

Zum Beispiel ist mehr als der Hälfte amerikanischer Facebook-NutzerInnen nicht bewusst, dass der Facebook-Newsfeed (die Nachrichten anderer Facebook-NutzerInnen, die man angezeigt bekommt) mittels prädiktiver Analytik algorithmisch kuratiert wird.[35] Ein umstrittenes realweltliches Experiment von Facebook-MitarbeiterInnen an Million Facebook-NutzerInnen hat gezeigt, dass eine Kuratierung des Feeds nach emotionaler Färbung signifikanten Einfluss auf die Emotionen und Handlungen von NutzerInnen hat.[36] Diese Mechanismen lassen sich auch dafür verwenden, das politische (Wahl-)Verhalten von Individuen zu beeinflussen.[37] Im Zusammenhang mit dem Cambridge-Analytica-Skandal wurde bekannt, dass NutzerInnen anhand ihrer Facebook-Aktivitäten mittels prädiktiver Analytik in psychologische Profilgruppen eingeteilt werden können, um ihnen individuell maßgeschneiderte politische Werbung anzuzeigen.[38] Kommentatorinnen gehen seit Längerem von einem erheblichen Einfluss dieser Technologie auf das politische Geschehen aus.[39]

Die Technologie KI-basierter Verhaltensprognostik ist meist in unsere (Daten-)Infrastrukturen (Social Media, Suchmaschinen, E-Mail-Anbieter, …) eingebaut und deshalb nicht direkt beobachtbar. Die „unsichtbare Hand", mit der diese Techniken unsere Handlungsfelder strukturieren, zeigt sich auch am Beispiel von KI im Human Resource Management.[40] Betrachten wir zum Beispiel ein auf maschinellem Lernen basierendes System, das anhand von Social-Media-Daten eine Reihe persönlicher Attribute (wie zum Beispiel „sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender"[41]) einzelner NutzerInnen prognostiziert und anhand dieser Informationen die BewerberInnen auf eine Job-Ausschreibung vorsortiert. Solche KIs sind *strukturelle* Agenten: So ist es erstens sehr wahrscheinlich, dass die BewerberInnen von der Existenz eines solchen Systems gar nichts erfahren, weil es ihnen nicht materiell entgegentritt. Zweitens sind die Grenzen dieses Systems unklar, da es sich bei einer solchen KI gar nicht um einen nur technischen, sondern um einen *sozio-technischen* Apparat handelt. Zahlreiche menschliche Akteure wirken daran mit, unter anderem die Millionen NutzerInnen der Social-Media-Plattform, die durch die bloße Benutzung dieses Dienstes das Datenmaterial (Erfahrungsbasis

35 Vgl. Eslami et al. (2015).
36 Vgl. Kramer et al. (2014).
37 Vgl. Bond et al. (2012); Tufekci (2014); sowie Persily (2017).
38 Vgl. Zittrain (2013) u. González (2017).
39 Vgl. Tufekci (2015).
40 Vgl. O'Neil (2016), 105–140, u. Bogen (2019).
41 Kosinski et al. (2013).

W) erzeugen, das zum Training des Vorhersagemodells benötigt wird. Die Social-Media-Plattform, inklusive ihrer NutzerInnen, gehört deshalb zu dem „hybriden Mensch-Maschine-System", das im Ganzen jene KI bildet, die die BewerberInnen klassifiziert.[42] Auch aus diesem Grund ist KI hier keine Informationsverarbeitungsleistung, die allein in einem Rechenzentrum lokalisierbar ist, sondern eine technisch-strukturelle Konfiguration, an der wir alle beteiligt sind.

Strukturalität KI-basierter Prognosesysteme bedeutet deshalb zweierlei: Die KI tritt den von ihren Prognosen betroffenen Individuen nicht materiell gegenüber, sondern rahmt ihren Zugang zu Ressourcen und Möglichkeiten. Und die Grenzen des KI-Systems sind unscharf, weil zahlreiche Menschen durch ihren Gebrauch digitaler Dienste und Medien passiv zur Intelligenzleistung des Systems beitragen. Weit von der imaginären Figur humanoider Roboter entfernt zeigt sich hier eine KI-Technologie in distribuierten Datennetzen, die als Möglichkeitsbedingung für Kommunikations- und Interaktionsverhältnisse auftritt und deshalb nur schwer als ethischer Agent aufzulösen ist. Ihr Einsatzpunkt ist die indirekte Beeinflussung des ökonomischen Verhaltens von Marktteilnehmern, etwa durch individualisierte Werbung oder individualisierte Preisanpassungen, und die algorithmische Verwaltung von Kohorten und Bevölkerungen durch Ungleichbehandlung der einzelnen Individuen auf Grundlage prognostizierter Verhaltensweisen, Risiken oder Effizienzpotenziale (Predictive Policing, Credit Scoring, individualisierte Preisgestaltung im Versicherungsbereich, Hiring Algorithms). Ihr Rohmaterial sind die Daten, die NutzerInnen täglich durch die Benutzung digitaler Dienste generieren.

2.3 Kausalität vs. Korrelation: Performativität KI-basierter Prognosesysteme

KI-basierte Prognosesysteme beruhen auf statistischen Vorhersagemodellen. Die Schlussweise solcher Verfahren ist grundsätzlich induktiv; sie beruht auf gemessenen Korrelationen und nicht auf bewiesenen Kausalzusammenhängen.[43] Das damit produzierte Wissen ist somit kein sicheres Wissen, sondern unterliegt der Möglichkeit induktiver Fehlschlüsse. Das sind Situationen, in denen eine bestimmte Verallgemeinerung auf empirischer Grundlage von *n* Fällen bei Anwendung auf einen *n+1*-ten Fall nicht mehr zutrifft.

42 Mühlhoff (2019a).
43 Vgl. Hildebrandt (2011) u. Mittelstadt et al. (2016).

Diese Gefahr falscher Verallgemeinerungen ist in der Praxis vor allem relevant, wenn es um *systematische* Biases in der Datengrundlage geht, weniger jedoch hinsichtlich abweichender Einzelfälle der betrachteten Kohorte. Denn die Datengrundlage W ist in vielen realen Einsatzbereichen von prädiktiver Analytik so groß (n im Bereich von Millionen Fällen), dass die abgeleiteten Korrelationen durch einzelne neue Beobachtungen (abweichender Fall $n+1$) kaum beeinflusst werden. Werden jedoch nur Daten bestimmter Populationen erhoben oder sind in der Stichprobe nicht alle Bevölkerungsgruppen gleichermaßen repräsentiert, dann kommt es zu Biases und Diskriminierung, sobald die statistischen Inferenzen auf Grundlage dieser beschränkten Stichprobe auf eine weiter gefasste Grundgesamtheit übertragen werden. Hierbei haben wir es dann mit dem Typus Bias zu tun, der bereits in den Trainingsdaten angelegt ist.[44]

Ein verwandtes Problem ist fundamentaler, jedoch weniger diskutiert: Selbst wenn eine gemessene Korrelation auf sehr vielen Datenpunkten basiert und deshalb von hoher statistischer Signifikanz und als „beinahe" kausales Wissen handhabbar ist, stellt sie eine statistische Erkenntnis *mit Bezug auf eine Gesamtkohorte* dar.[45] Statistische Prognosesysteme jedoch transferieren solche (stets gesamtheitsbezogenen) Inferenzen auf Einzelfälle und vereindeutigen damit das Resultat: Im allgemeinen ist das Ergebnis $A_i = P_W(D_i)$ einer prädiktiven Analyse für das Input-Datum D_i kein *eindeutiger* Ausgabewert, sondern eine Wahrscheinlichkeitsverteilung über verschiedene mögliche Werte – zum Beispiel: Mit 55 % Wahrscheinlichkeit fällt das Individuum i in Kategorie A, mit 30 % in Kategorie B, mit 15 % in Kategorie C, … Sobald auf die Prognose jedoch eine Handlung (Entscheidung) folgt, muss die statistische Unschärfe der Prädiktion aufgelöst werden, indem der Fall entweder als Kategorie A, B oder C behandelt wird.

Statistische und KI-basierte Prognosesysteme nehmen in diesem Sinne einen Sprung von der populationsbezogenen Inferenz zur Prognose über einen Einzelfall vor. Das ist noch etwas anders als falsche Verallgemeinerung: Der Fehler entsteht hier nicht durch Bildung einer Allgemeinaussage anhand endlicher empirischer Evidenz, sondern durch Bildung einer Einzelfallaussage anhand statistischen Wissens über eine Kohorte. Bekannt ist eigentlich nur ein Wissen über die Kohorte, die den Trainingsdaten zugrunde liegt: Individuen mit der Datensignatur D_i fallen innerhalb dieser Kohorte in 60 % der Fälle in die Gruppe A, in 35 % der Fälle in die Gruppe B und in 15 % der Fälle in die Gruppe C. Das ist ein belastbares statistisches Wissen, keine falsche Verallgemeinerung. Doch wenn sich das Prognosesystem für A, B oder C entscheiden muss und real-weltliche

44 Vgl. Friedman/Nissenbaum (1996) sowie Mittelstadt et al. (2016).
45 Vgl. auch ebd., 5.

Auswirkungen auf die Behandlung des Individuums *i* hat, dann tritt die Gefahr einer Falschbehandlung auf.

Man kann hier von der *Performativität statistischer Prognosesysteme* sprechen, weil die vereindeutigende Entscheidung durch Ungleichbehandlung in die Welt eingreift und die betreffende Eigenschaft des betroffenen Individuums damit hervorbringen kann.[46] Ist die geschätzte Bonität eines Individuums *i* zum Beispiel unscharf zwischen den Kategorien „mittel" und „schlecht" verteilt, dann kann es passieren, dass die Person durch die Vereindeutigung als „schlecht" behandelt wird und höhere Kreditzinsen für sie veranschlagt werden, was ihre finanzielle Belastung steigert und ihre schlechte Bonität somit zusätzlich produziert. Cathy O'Neil weist solche Feedbackeffekte als systematische Eigenschaft des realen Einsatzes prädiktiver Analytik aus, die dadurch die Realität, die sie prognostizieren, hervorbringen oder zumindest stabilisieren können.[47]

Der Sprung von statistischer Inferenz (die immer ein Wissen über Kohorten ist) zu handlungsleitenden Prognosen über Einzelfälle markiert auch den begrifflichen Unterschied zwischen den eingangs definierten Begriffen „KI-basiertes Prognosesystem" und „prädiktive Analyse". Ist die prädiktive Analyse bloß die mathematische Funktion P_w, der man zunächst einen rein deskriptiven Status zuschreiben kann, bezeichnet das KI-basierte Prognosesystem den gesamten soziotechnischen Zusammenhang, in dem die prädiktive Analyse verwendet und mit realen Auswirkungen verschaltet wird. Dazu gehört die Einbindung des Verfahrens in reale Entscheidungskontexte, in denen Inputdaten D_i für den Algorithmus gewonnen und die Ausgabe A_i handlungswirksam in einen realen Kontext zurückgespielt wird. In dieser Wechselwirkung mit einer Umgebung liegt ein gewichtiger ethischer Problembereich prädiktiver Systeme, der aus der Umwandlung (kohortenbezogener) statistischer Inferenz in vereindeutigende Einzelfallprognosen resultiert.

3 KI-Ethik und Sozialtheorie

Maschinelles Lernen im Bereich prädiktiver Analytik (PA) und automatisierter Entscheidungsfindung bildet einen der aktuell größten und relevantesten Anwendungsbereiche von KI-Technologie und zeigt bereits jetzt flächendeckende gesellschaftliche Auswirkungen in einer Vielzahl von Lebensbereichen. Für eine

46 Vgl. zur Performativität von Daten Matzner (2016).
47 Vgl. O'Neil (2016), 148–155.

ethische Thematisierung spielt es nun eine große Rolle, dass diese Technologien untrennbar mit den Geschäftspraktiken der modernen IT-Industrie verbunden sind. Das liegt an mindestens zwei Faktoren: Erstens ist PA überhaupt nur im Kontext digitaler „Business-to-Customer"-Operationen (B2C) entstanden. Denn hierbei handelt es sich um ein zentrales Instrument jener Spielart des post-industriellen Kapitalismus, die auf individuell zugeschnittene Angebote und Dienstleistungen setzt.[48] Prognosen über zukünftiges Verhalten oder Reaktionsweisen der MarktteilnehmerInnen spielen dabei eine große Rolle, um Risiken zu kontrollieren und zukünftige Potenziale abzuschätzen. Zweitens werden PAs auf großen Mengen von Trainingsdaten trainiert, die oft als Nebenprodukt durch die alltägliche Benutzung von digitalen Diensten und *Internet-of-Things*-Anwendungen entstehen. Nutzungsdaten, Ortsinformationen, Körperdaten, Sensordaten, Zahlungsdaten und Transaktionshistorien werden von NutzerInnen täglich generiert – und werden als Erfahrungswissensbestände für Prognosesysteme aller Art verwendet.[49] Die zunehmende Kolonisierung aller Bereiche des privaten und öffentlichen Lebens durch vernetzte Digitaltechnik, die mit dem Technologietrend des „Ubiquitous Computing" zusammenhängt,[50] führt zu einem stetigen Anwachsen des alltäglich erzeugten Datenvolumens und fördert damit die Bereichserweiterung prädiktiver Analytik.[51]

Um diese Zusammenhänge mit analytischer Tiefe zu behandeln, ist es entscheidend, dass Ethik der KI umfassend sozialtheoretisch und polit-ökonomisch informiert ist. Das bedeutet, in die ethische Betrachtung eine Analyse der Machtmuster, Subjektivierungsformen,[52] Kapitalakkumulationsweisen und lokalen wie globalen Ungleichheiten einzubeziehen, die mit KI-Technologie in unlösbarem Zusammenhang stehen.[53] Insbesondere die Verzahnung der KI-Technologie mit den Wertschöpfungsformen und privatisierten Märkten des „digitalen Kapitalismus"[54] ist dafür relevant. Denn wir haben es im Fall der KI-Technologie mit einer globalen industriellen Entwicklungsdynamik zu tun, die durch erhebliche Kapitalinteressen gedeckt ist und potenziell unsere Kommunikations-, Interaktions- und Wirtschaftsformen in zahlreichen Bereichen neu organisiert. Diejenigen KI-Technologien, die heute greifbare Auswirkungen auf Gesellschaft und Politik haben, sind nicht von der IT-Industrie, ihrer Innovationskultur, ihren

48 Vgl. Daum (2019).
49 Vgl. Wachter (2018).
50 Vgl. Weiser (1991) u. Kaerlein (2018).
51 Vgl. Mühlhoff (2019a).
52 Vgl. ders. (2018).
53 Vgl. Daum (2019).
54 Staab (2019).

Datennetzwerken und ihrer quasi-infrastrukturellen Verwobenheit mit den Alltagspraktiken von Millionen NutzerInnen ablösbar.

Für eine ethische Thematisierung von KI, die dem Problem der prädiktiven Analytik gewachsen ist, folgt hieraus zweierlei: Diskriminierung, Bias und Opazität prognostischer Verfahren sind grundsätzlich mit sozio-ökonomischen Machtgefällen, globaler Ungleichheit und kompetitiven Geschäftspraktiken verschränkt. Wie Safiya Umoja Noble in ihrer Studie über rassistische Biases der Google-Suchmaschine argumentiert, handelt es sich bei diesen ungerechten Auswirkungen nicht etwa bloß um einen „Fehler" im System, den wohlmeinende IngenieurInnen irgendwann finden und ausmerzen werden.[55] Hier repliziert sich vielmehr eine sozio-ökonomisch tief verankerte Machtkonstellation globaler Kapitalinteressen.

Denunziation diskriminierender Entscheidungen oder Prognosen, Entlarvung von Biases und der Ruf nach Transparenz der algorithmischen Routinen sind längst Gemeinplätze in der Algorithmenethik und den zahlreichen Ethik-Richtlinien, die private, politische, zivilgesellschaftliche und wissenschaftliche Akteure herausgebracht haben.[56] Doch diese Instrumente verfehlen eine verlässliche Wirkungskraft, solange nicht die tiefen Verflechtungen der benannten Missstände mit wirtschaftlichen Profit-, Wettbewerbs- und Akkumulationsprinzipien erkannt werden. Im Kontext dieser Strukturen erfüllt die KI-Technologie eine Vermittlungs- und Katalysefunktion: Prädiktive Analysen werden schlechterdings *dafür gebaut, zu diskriminieren*, wenn damit im ursprünglichen Wortsinn die laterale Unterscheidung und Abgrenzung von Fällen gemeint ist. Das Bekenntnis der beteiligten Firmen zu „ethischer" KI-Technologie ist deshalb ein Täuschungsmanöver. Setzt man KI-basierte Prognostik zum Beispiel für individuelle Versicherungsbepreisung ein (wer ungesünder lebt, zahlt mehr für die Krankenversicherung[57]), dann hat man sich bereits implizit dafür entschieden, das Solidarmodell des Risiko-Poolings aufzugeben, ganz gleich, ob der prädiktive Mechanismus Biases zeigt oder nicht. Eine Ethik, die nur Biases ausmerzen möchte, will auf inkrementelle Verbesserungen eines *an sich* problematischen Systems hinaus. Ob wir diese Systeme zur Beurteilung von Menschen *überhaupt* einsetzen möchten, ist das zentrale ethische Problem, das jedoch tiefer liegt als die Begriffe Bias, Diskriminierung, Opazität der Algorithmen-Ethik etc. zu blicken gestatten. Ethik der KI muss die Frage stellen, wie wir uns gegenseitig behandeln möchten: Als individuelle Risikofaktoren, die es am äußerlichen Verhalten zu erkennen und

55 Noble (2017).
56 Vgl. Mittelstadt et al. (2016); Mittelstadt (2019); sowie Hagendorff (2019).
57 Vgl. O'Neil (2016), 174–175.

zu managen gilt, oder als Solidargemeinschaft, die auf verinnerlichten Werten basiert und in der Risiken verteilt und gemeinsam geschultert werden.

Zweitens zeigt sich in der Verflochtenheit prädiktiver Analytik mit der digitalen Vernetzung, dass die NutzerInnen digitaler Dienste als tägliche LieferantInnen von Trainingsdaten eine zentrale Rolle in diesem System spielen.[58] Eine wirkungsvolle ethische Thematisierung muss deshalb auch die Prägung von Nutzungsgewohnheiten, digitalen Subjektivitäten und Körper-Technik-Verhältnissen durch digitale Artefakte und Kulturen addressieren. Beim Studium der Geschichte des kommerziellen User Experience Design etwa zeigt sich, dass Entwicklungen hin zur intuitiven Bedienbarkeit graphischer Oberflächen eng mit der Erfolgsgeschichte des maschinellen Lernens verknüpft sind.[59] Erst seitdem durch das interaktive Web 2.0, durch die Ausbreitung von Web Analytics (flächendeckendes Tracking zur Design-Optimierung von Webseiten) und die Verbreitung von Social Media umfangreiche Infrastrukturen für behaviorale Untersuchungen im Internet entstanden sind, wurden bei den Plattformunternehmen genügend große Vorräte an Trainingsdaten angesammelt, um in kommerziellen Anwendungen des maschinellen Lernens signifikante Erfolge verzeichnen zu können.[60] Die Hervorbringung der Mensch-Technik-Relation als ein Teilaspekt von Subjektivität ist deshalb ein wichtiges Analysefeld für Ethik der KI, denn die heute relevante KI, die stets auf großen Datenmengen basiert, wäre nicht ohne die freiwillige und unbemerkte Mitarbeit einer Mehrheit der Gesellschaftsmitglieder durch tägliche Datenproduktion denkbar.

3.1 Biopolitische Wende des digitalen Kapitalismus

Wirkungsvolle, sozialtheoretisch informierte Ethik der KI ist auch deshalb ein besonders drängendes Problem, weil wir uns an einem entscheidenden Punkt in der Entwicklung der vernetzten Digitaltechnik im Zusammenhang mit PA befinden: Das Anwendungsfeld von PA ist aktuell im Begriff, den Bereich der B2C-Operationen zu überschreiten und zunehmend die Relationen zwischen Staat und StaatsbürgerInnen („Government to Citizen", G2C) zu erfassen. Profilbildung, Risiko-Scoring und automatisierte Entscheidungsfindung werden nicht mehr nur für gezielte Werbung, differenzielle Preisgestaltung und andere Strategien zur

58 Vgl. Mühlhoff (2019a).
59 Vgl. ders. (2018).
60 Vgl. ebd. u. ders. (2019a u. 2019b).

Manipulation *des wirtschaftlichen Verhaltens von Marktteilnehmern* eingesetzt. Vielmehr werden diese Techniken nun dafür verwendet, die Beziehung des Einzelnen zum Staat, einschließlich der wohlfahrtsstaatlichen Institutionen, der Bildungs- und Gesundheitseinrichtungen, des Sicherheitsapparats und der Politik zu regeln. Prädiktive Analytik wird zunehmend als Werkzeug für *algorithmisches Bevölkerungsmanagement* eingesetzt: das zeigt sich etwa im Feld der prädiktiven Polizeiarbeit (Predictive Policing[61]), wenn ML im Gesundheits- und Versicherungssystem zum Einsatz kommt[62], in der Kriminologie zur Vorhersage von Rückfallwahrscheinlichkeiten[63], im Bildungsbereich[64], auf dem Arbeitsmarkt[65], im Jugendschutz[66] und im psychologischen Targeting von politischer Werbung[67].

Virginia Eubanks hat den in diesem Zusammenhang passenden Begriff der „Automatisierung von Ungleichheit" ins Spiel gebracht:[68] Werden statistische und KI-basierte Verfahren der Profilbildung, Risikobewertung und individualisierten Ansprache (Targeting) zur Regulierung des individuellen Zugriffs auf staatliche Ressourcen oder zur differenziellen Belegung mit Beschränkungen (Sicherheit, Polizeiarbeit, Pandemiebekämpfung, Jugendschutz) verwendet, dann wird – teils entlang neuer, teils entlang bestehender Strukturen – ein soziales Gefälle produziert. Auf undurchsichtige Weise werden Individuen in Bezug auf den Zugang zu Ressourcen, Chancen und die Durchsetzung ihrer Rechte unterschiedlich behandelt. Aus individualisierter Werbung wird so ein neues Prinzip der *sozialen Selektion* mit dem Effekt, dass sich die Gesellschaft in unsichtbare soziale Klassen organisiert, z. B. in solche, die angeblich ein Sicherheits- oder Gesundheitsrisiko darstellen, vorrangigen Zugang zu knappen medizinischen Ressourcen erhalten, aufgrund ihres Lernverhaltens in der Schule oder an der Universität angeblich für bestimmte Berufe geeignet sind oder eher Opfer häuslicher Gewalt werden und deshalb präventiv vom Jugendschutz überwacht werden sollten.[69] Es ist nicht nötig, auf das chinesische „Social Credit System" zu verweisen, um die Diagnose zu begründen, dass sich Prädiktive Analytik und KI aktuell von der Kontrolle des Marktverhaltens zur Kontrolle der Beziehungen zu staatlichen Institutionen verlagert und eine neue technologische Ära der „Biopolitik" vorbereitet: Damit ist

61 Vgl. Crawford/Schultz (2014).
62 Vgl. Prainsack (2020).
63 Vgl. Hao (2019).
64 Vgl. Baker/Inventado (2014).
65 Vgl. O'Neil (2016).
66 Vgl. Eubanks (2016).
67 Vgl. Persily (2017) sowie Tufekci (2014 u. 2015).
68 Eubanks (2017).
69 Vgl. ebd., 127 ff.

ein Mechanismus der algorithmischen Bevölkerungsverwaltung gemeint, der tief in die biologischen, ökonomischen und sozialen Prozesse integriert ist und sich zu einem Macht- und Kontrollapparat zusammensetzt.[70]

Zwei weitere Kennzeichen dieser biopolitischen Bereichserweiterung von PA stellen große Herausforderungen an eine Ethik der KI: Erstens wird sie nicht, wie man meinen könnte, durch einen Transfer von KI-Methoden und -Techniken vom privaten in den öffentlichen Sektor erreicht. Zu beobachten ist vielmehr eine tendenzielle Integration staatlicher Institutionen in private Plattformen, Datennetzwerke und Mensch-Maschine-Schnittstellen.[71] Dies liegt daran, dass ML-Technologie mit den Wertschöpfungsformen und privaten Infrastrukturen des digitalen Kapitalismus verstrickt ist (siehe oben), die anderswo und getrennt von den Wirtschaftsakteuren nicht einfach repliziert werden können. Die Verwendung der KI-Techniken im G2C-Bereich geht deshalb mit einer Privatisierung von öffentlicher Fürsorgeinfrastruktur einher. Die biopolitische Verschiebung der ML-Technologie ist somit eine biopolitische Verschiebung *des digitalen Kapitalismus insgesamt*, die staatliche Institutionen und Kommunikation zwischen Staat und StaatsbürgerInnen schrittweise in die private Infrastruktur und ihre Wertschöpfungslogiken integriert.

Hier kommt nun zweitens hinzu, dass die biopolitische Verschiebung des Kapitalismus auch eine Verschiebung *der Biopolitik* ist – mit einer bedenkenswerten ethischen Konsequenz. Prädiktive Analytik, Behavioral Targeting und Risk-Scoring lenken das Interesse weg von der Vergangenheit einer Person hin zu ihrer prognostizierten Zukunft. Eine im Wesentlichen probabilistische Form der Argumentation ist daran beteiligt, wie BürgerInnen durch PA behandelt und Bevölkerungen verwaltet werden. Natürlich hat sich Biopolitik, wie Michel Foucault argumentiert, von Beginn an mit statistischen Erkenntnissen in Bezug auf die Verwaltung von Bevölkerungen beschäftigt.[72] Dies zeigt sich zum Beispiel in den öffentlichen Gesundheitsprogrammen, wenn etwa eine Impfung aller gegen eine bestimmte Krankheit angeordnet und dabei in Kauf genommen wird, dass einige wenige aufgrund der Nebenwirkungen sterben werden, um das Leben

70 Vgl. Foucault (2006a u. 2006b). Der Begriff wurde später vor allem in den Gouvernementalitätsstudien aufgegriffen, vgl. Bröckling et al. (2000).

71 Z. B. fördern Gesundheitsbehörden in der Corona-Pandemie die Entwicklung von Smartphone-Apps zur erleichterten Infektionskettenverfolgung, anstatt ihre eigene (mutmaßlich ineffiziente) Infrastruktur zur Kontaktverfolgung zu verwenden. Diese Apps laufen auf kommerziellen Plattformen und die Diskussion um die Gestaltung der Apps zeigt deutlich, dass die privaten Akteure Apple und Google hier den Rahmen setzen.

72 Zur Biopolitik der medizinischen Institutionen und zum statistischen Denken in der Verwaltung von Bevölkerungen vgl. Foucault (2006a).

vieler anderer zu schützen. Während sich die klassische Statistik jedoch auf Durchschnittswerte und Kohorten im Allgemeinen bezieht, richtet sich prädiktive Analytik *auf das Individuum* – jedoch nicht per se, sondern insofern es sich innerhalb großer Kohorten vergleichen und einordnen lässt. Die Idee besteht nicht mehr darin, alle Individuen mit Blick auf einen bestimmten Durchschnittseffekt (der für eine zufällige und unbekannte Minderheit nachteilig und für eine Mehrheit vorteilhaft sein wird) *gleich* zu behandeln, sondern sie alle *unterschiedlich* zu behandeln, je nach ihrer individuellen Risikoeinschätzung innerhalb der Kohorte. Da diese Risikoeinschätzungen durch die Übertragung statistischer Inferenzen (bei denen es immer um Gruppen und Kohorten geht) auf den Einzelfall operieren, wird die Statistik in diesem Verarbeitungsschritt zur Prognostik. Die Verwendung individueller Vorhersagen über zukünftiges Verhalten, Risiken, Potentiale und Verwundbarkeiten von Personen anstelle von aggregierten Durchschnittswerten einer Bevölkerung als Grundlage für die Behandlung von BürgerInnen stellt einen Paradigmenwechsel in der Biopolitik dar und markiert den Übergang von einer Ethik des Solidarmodells (alle werden gleich behandelt, das Risiko wird geteilt und im Durchschnitt gemindert) zu einer Ethik der sozio-ökonomischen Selektion – jede/r wird behandelt, „wie er/sie es verdient".

4 Schluss: KI-Ethik und kritische Praxis

Weil der Einsatz von KI-Technologie im Begriff ist, unsere sozialen, gesellschaftlichen, politischen und wirtschaftlichen Beziehungen grundlegend zu strukturieren, muss sich eine wirkungsvolle Ethik der KI zur einer starken Rolle ethischer Diskurse für eine kollektive Bildungs-, Reflexions- und Kritikpraxis bekennen. Hierbei ist mit Bildung, Reflexion und Kritik die reflexive Arbeit einer politischen Gemeinschaft an sich selbst gemeint. In diese Arbeit müssen alle eingebunden werden und sie muss im Kontext eines gesellschaftlichen Zusammenhalts stehen, so wie etwa die Frage der Aufklärung eine der Aktivierung jedes Einzelnen im Geist einer kollektiven historischen Bewusstwerdung war.[73] Ethik der KI muss die beteiligten und betroffenen Subjekte *ansprechen*, sie als ethische Agenten aktivieren und in die Pflicht nehmen, um wirklich in Veränderungen kanalisiert zu werden und nicht einen Urteilsstandpunkt zu beanspruchen, der dem Geschehen enthoben ist.

73 Vgl. Foucaults (2007) Referat über das Kant'sche Verständnis von Aufklärung.

Dabei ist an mindestens drei Gruppen von Beteiligten und Betroffenen zu denken: (1) die EntwicklerInnen, ProfiteurInnen und „Shareholder" der neuen Technologien; (2) die EntscheiderInnen, PolitikerInnen und zivilgesellschaftlichen Akteure, die professionell mit ihrer Regulierung zu tun haben; (3) die NutzerInnen, Betroffenen und „Stakeholder". Es ist besonders markant, dass gerade die Gruppe der NutzerInnen – d. h. die Frage der *Benutzung* etwa digitaler Dienste – bisher fast gar nicht im Fokus ethischer Thematisierungen der KI steht, insbesondere nicht in der Weise, dass den NutzerInnen qua ihrer Benutzungsweise digitaler Dienste eine Mitverantwortung zugesprochen wird.[74] Wie schon beschrieben spielen die NutzerInnen, ihre Gewohnheiten, Vorlieben und Subjektivierungsformen gerade bei datenbasierten Anwendungen des maschinellen Lernens eine wesentliche Rolle; ohne die von uns allen generierten Daten gäbe es diese KI-Technologien nicht in ihrer heutigen Form. Tatsächlich sind die NutzerInnen, zum Beispiel der Google-Suchmaschine, selbst als ein Baustein des Systems zu betrachten, das die Intelligenzleistung dieses Apparats vollbringt.

Es liegt in Bezug auf datenbasierte KI-Technologien eine beinahe paradoxale, transzendentale Bedingungsrelation zwischen der Technologie und der Subjektivität ihrer NutzerInnen vor: Einerseits werden diese Technologien überhaupt erst möglich durch die Benutzung, das heißt durch die Gewohnheiten, Prägungen und sozialen Interaktionsformen, die zu ihrer unweigerlichen Benutzung führen. Andererseits geht es hier um Systeme, die die technologischen Rahmenbedingungen unseres Denkens und Wissens, unserer Politik, Ökonomie und Öffentlichkeit sowie unserer sozialen Interaktionen und Kulturaktivitäten bestimmen. Diese Technologien sind Konstitutionsbedingungen unseres Denkens, Fühlens und Handelns und zugleich hängen sie existenziell von unserer Mitarbeit ab.

Wenn Ethik der KI auf eine aktive und bewusste Gestaltung unseres Verhältnisses zu diesen Techniken hinauslaufen soll, dann ist es unerlässlich, diese technologische Bedingtheit unserer Wissens-, Politik- und Sozialkultur zu studieren und zu kritisieren.[75] Dies impliziert die Analyse der Art und Weise, wie Macht- und Herrschaftsstrukturen in technologische Artefakte eingeschrieben sind. Die ethischen Fragen der KI laufen deshalb nicht auf Fragen der Form „Was soll ich tun?" hinaus, die suggerieren, dass es bloß verantwortungsvoll zwischen *gleichermaßen vor Augen stehenden* Handlungsoptionen abzuwägen gälte. Die Handlungsoptionen angesichts einer Technologie, die unseren Weltzugang so wesentlich prägt, können *nicht* so leicht überblickt werden. Ethik der KI muss

74 Vgl. die Idee einer „distribuierten Ethik der KI" in Wong/Simon (2020) und zur Rolle der NutzerInnen Mühlhoff (2018).

75 Breljak/Mühlhoff (2019).

sich mit der transzendentalen, aufklärerischen Frage unserer technologischen Konstitution befassen, um das Problem in seiner vollen Komplexität zu erfassen.

Aus diesem Grunde benötigt Ethik der KI eine *kritische Theorie*, wenn kritische Theorie bedeutet, die technologischen Bedingungsfaktoren sowie die darin eingeschriebenen Machtstrukturen und Herrschaftsverhältnisse hervorzuarbeiten, die unser Denken, Fühlen und Handeln konstituieren. Zugleich jedoch wäre eine solche kritische Theorie ohne Ethik nur eine Theorie, denn der Ethik wiederum kommt die Aufgabe zu, die unter den bestehenden Bedingungen für jede/n Einzelne/n ergreifbaren Möglichkeiten zur Veränderung aufzuweisen und in praktisches Handeln zu überführen. So müssen kritische Theorie des digitalen Kapitalismus und Ethik der Künstlichen Intelligenz eine „dynamische Einheit" im Sinne Max Horkheimers bilden: Die Darstellung der Missstände und ihre ethische Bewertung müssen als „stimulierender, verändernder Faktor" in der Gesellschaft wirken.[76] Ohne tiefe Grabungen, die die technologischen Konstitutionsbedingungen unserer selbst ständig neu zutage fördern und problematisieren, würde Ethik der KI die Tragweite ihres Gegenstands und die nötige Wirkungskraft verfehlen.

Literatur

Anderson, M., u. Anderson, S. L. (Hg.) (2011), Machine ethics, New York.

Baker, R. S., u. Inventado, P. S. (2014), Educational Data Mining and Learning Analytics, in: Larusson, J. A., u. White, B. (Hg.), Learning Analytics: From Research to Practice, New York, 61–75.

Bogen, M. (2019), All the Ways Hiring Algorithms Can Introduce Bias, in: Harvard Business Review, URL: https://hbr.org/2019/05/all-the-ways-hiring-algorithms-can-introduce-bias (27.5.2020).

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., u. Fowler, J. H. (2012), A 61-million-person experiment in social influence and political mobilization, in: Nature 489.7415, 295–298.

Bostrom, N., u. Yudkowsky, E. (2014), The ethics of artificial intelligence, in: Frankish, K., u. Ramsey, W. M. (Hg.), The Cambridge Handbook of Artificial Intelligence, Cambridge, 316–334.

Breljak, A., u. Mühlhoff, R. (2019), Was ist Sozialtheorie der Digitalen Gesellschaft? – Einleitung, in: Slaby, J. (Hg.), Affekt Macht Netz: Auf dem Weg zu einer Sozialtheorie der digitalen Gesellschaft, Bielefeld, 7–34.

Bröckling, U., Krasmann, S., u. Lemke, T. (Hg.) (2000), Gouvernementalität der Gegenwart: Studien zur Ökonomisierung des Sozialen, Frankfurt am Main.

76 Horkheimer (1992), 232.

Crawford, K., u. Schultz, J. (2014), Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms, in: Boston College Law Review 55.1, 93–128.
Daum, T. (2019), Die Künstliche Intelligenz des Kapitals, Hamburg.
Dignum, V. (2019), Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way, Cham, ZG.
Duhigg, C. (2012), How Companies Learn Your Secrets, in: The New York Times, URL: https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html (27.5.2020).
Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., u. Sandvig, C. (2015), „I always assumed that I wasn't really that close to [her]" Reasoning about Invisible Algorithms in News Feeds, in: Proceedings of the 33rd annual ACM conference on human factors in computing systems, 153–162.
Etzioni, A., u. Etzioni, O. (2017), Incorporating Ethics into Artificial Intelligence, in: The Journal of Ethics 21.4, 403–418.
Eubanks, V. (2017), Automating inequality: how high-tech tools profile, police, and punish the poor, New York.
Foucault, M. (2006a), Sicherheit, Territorium, Bevölkerung: Geschichte der Gouvernementalität I, Frankfurt am Main.
Foucault, M. (2006b), Die Geburt der Biopolitik: Geschichte der Gouvernementalität II, Frankfurt am Main.
Foucault, M. (2007), Was ist Aufklärung? [1984], in: Ästhetik der Existenz: Schriften zur Lebenskunst, Frankfurt am Main, 171–190.
Friedman, B., u. Nissenbaum, H. (1996), Bias in computer systems, in: ACM Transactions on Information Systems 14.3, 330–347.
Goggin, B. (2019), Inside Facebook's suicide algorithm: Here's how the company uses artificial intelligence to predict your mental state from your posts, in: Business Insider, URL: https://www.businessinsider.com/facebook-is-using-ai-to-try-to-predict-if-youre-suicidal-2018-12 (27.5.2020).
González, R. J. (2017), Hacking the citizenry?: Personality profiling, „big data" and the election of Donald Trump, in: Anthropology Today 33.3, 9–12.
Grindrod, P. (2014), Mathematical underpinnings of analytics: theory and applications, Oxford.
Grunwald, A. (Hg.) (2013), Handbuch Technikethik, Stuttgart.
Hagendorff, T. (2020), The Ethics of AI Ethics – An Evaluation of Guidelines, in: Minds and Machines 30, 99–120.
Hao, K. (2019), AI is sending people to jail – and getting it wrong, in: MIT Technology Review, URL: https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/ (27.5.2020).
Hildebrandt, M. (2011), Who Needs Stories if You Can Get the Data? ISPs in the Era of Big Number Crunching, in: Philosophy & Technology 24.4, 371–390.
Horkheimer, M. (1992), Traditionelle und kritische Theorie [1937], in: Traditionelle und kritische Theorie: fünf Aufsätze, Frankfurt am Main, 205–259.
Hurley, M., u. Adebayo, J. (2017), Credit scoring in the era of big data, in: Yale Journal of Law and Technology 18.1, 5.
Jobin, A., Ienca, M., u. Vayena, E. (2019), The global landscape of AI ethics guidelines, in: Nature Machine Intelligence 1.9, 389–399.
Kaerlein, T. (2018), Smartphones als digitale Nahkörpertechnologien: Zur Kybernetisierung des Alltags, Bielefeld.

Kosinski, M., Stillwell, D., u. Graepel, T. (2013), Private traits and attributes are predictable from digital records of human behavior, in: Proceedings of the National Academy of Sciences 110.15, 5802–5805.
Kramer, A. D. I., Guillory, J. E., u. Hancock, J. T. (2014), Experimental evidence of massive-scale emotional contagion through social networks, in: Proceedings of the National Academy of Sciences 111.24, 8788–8790.
Lippert, J. (2014), ZestFinance issues small, high-rate loans, uses big data to weed out deadbeats, in: Washington Post, URL: 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 (27.5.2020)
Matzner, T. (2019), Autonomy Trolleys und andere Probleme: Konfigurationen künstlicher Intelligenz in ethischen Debatten über selbstfahrende Kraftfahrzeuge, in: Zeitschrift für Medienwissenschaft 21.2, 46–55.
Matzner, T. (2016), Beyond data as representation: The performativity of Big Data in surveillance, in: Surveillance & Society 14.2, 197–210.
Merchant, R. M., Asch, D. A., Crutchley, P., Ungar, L. H., Guntuku, S. C., Eichstaedt, J. C., Hill, S., Padrez, K., Smith, R. J., u. Schwartz, H. A. (2019), in: Evaluating the predictability of medical conditions from social media posts, in: PLOS ONE 14.6, e0215476.
Misselhorn, C. (2018), Grundfragen der Maschinenethik, 4. Aufl., Stuttgart.
Mittelstadt, B. (2017), From Individual to Group Privacy in Big Data Analytics, in: Philosophy & Technology 30.4, 475–494.
Mittelstadt, B. (2019), Principles alone cannot guarantee ethical AI, in: Nature Machine Intelligence 1.11, 501–507.
Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S., u. Floridi, L. (2016), The Ethics of Algorithms: Mapping the Debate, in: Big Data and Society 3.2.
Mühlhoff, R. (2018), Digitale Entmündigung und User Experience Design: Wie digitale Geräte uns nudgen, tracken und zur Unwissenheit erziehen, in: Leviathan – Journal of Social Sciences 46.4, 551–574.
Mühlhoff, R. (2019a), Human-Aided Artificial Intelligence: Or, How to Run Large Computations in Human Brains? Towards a Media Sociology of Machine Learning, in: New Media & Society OnlineFirst Nov. 2019.
Mühlhoff, R. (2019b), Big Data is Watching You. Digitale Entmündigung am Beispiel von Facebook und Google, in: Slaby, J. (Hg.), Affekt Macht Netz: Auf dem Weg zu einer Sozialtheorie der digitalen Gesellschaft, Bielefeld, 81–107.
Noble, S. U. (2018), Algorithms of oppression: how search engines reinforce racism, New York.
O'Dwyer, R. (2018), Are You Creditworthy? The Algorithm Will Decide, in: Undark Magazine, URL: https://undark.org/2018/05/07/algorithmic-credit-scoring-machine-learning/ (8.10.2020).
O'Neil, C. (2016), Weapons of math destruction: how big data increases inequality and threatens democracy, New York.
Owen, R., Bessant, J. R., u. Heintz, M. (Hg.) (2013), Responsible innovation, Chichester.
Persily, N. (2017), Can Democracy Survive the Internet?, in: Journal of Democracy 28.2, 63–76.
Prainsack, B. (2020), The value of healthcare data: to nudge, or not?, in: Policy Studies 41.5, 547–562.
Saslow, K., u. Lorenz, P. (2019), Artificial Intelligence Needs Human Rights, Berlin.
Staab, P. (2019), Digitaler Kapitalismus: Markt und Herrschaft in der Ökonomie der Unknappheit, Berlin.

Tufekci, Z. (2015), Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency, in: Colorado Technology Law Journal 13, 203.
Tufekci, Z. (2014), Engineering the public: Big data, surveillance and computational politics, in: First Monday 19.7, DOI: doi: http://dx.doi.org/10.5210/fm.v19i7.4901 (8.10.2020).
Véliz, C. (2019), Three things digital ethics can learn from medical ethics, in: Nature Electronics 2.8, 316–318.
Wachter, S. (2018), Normative challenges of identification in the Internet of Things: Privacy, profiling, discrimination, and the GDPR, in: Computer Law & Security Review 34.3, 436–449.
Wagner, B. (2018), Ethics as an escape from regulation. From „ethics-washing" to ethics-shopping?, in: Bayamlioğlu, E., Baraliuc, I., Janssens, L., et al. (Hg.), Being Profiled: Cogitas Ergo Sum. 10 Years of 'Profiling the European Citizen', Amsterdam, 84–88, DOI: https://doi.org/10.2307/j.ctvhrd092.18.
Weiser, M. (1991), The computer for the 21st century, in: ACM SIGMOBILE mobile computing and communications review 3, 3–11.
Wong, P.-H., u. Simon, J. (2020), Thinking About ‚Ethics' in the Ethics of AI, in: Idees 48, URL: https://revistaidees.cat/en/thinking-about-ethics-in-the-ethics-of-ai/ (8.10.2020).
Zittrain, J. (2013), Engineering an election, in: Harvard Law Review Forum 127, 335.
Zwitter, A. (2014), Big Data ethics, in: Big Data & Society 1.2, 205395171455925.