10.6094/behemoth.2021.14.2.1057

BEHEMOTH A Journal on Civilisation
2021 Volume 14 Issue No. 2

Warum Ethikstandards nicht alles sind. Zu den herrschafts-konservierenden Effekten aktueller Digitalisierungskritik

Why Ethic Standards are not Enough, and how Current Critiques of Digitalization Preserve Power

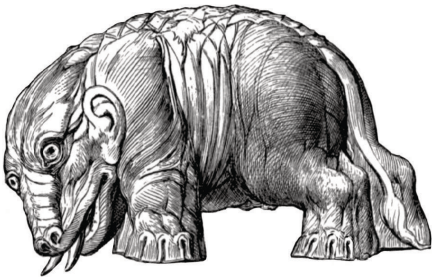
Bianca Prietl

Abstract

This article deals with *AI ethic guidelines and standards* as the currently dominant form in which society articulates critique of digital data technologies and searches for solutions for the respective upheavals. Based on a discourse-analytical reflection, it is argued that the premises underpinning the societal critique of digitalization have several conceptual limitations – especially when it comes to understanding and questioning social relations of power and the role that digital data technologies play in their reproduction. Against this backdrop, the currently dominant form of societal critique of digitalization is described as essentially preserving power relations. Therefore, it is pleaded for strengthening rationality- and power-critical perspectives in the debates on digitalization and its challenges.

Keywords, dt.: Digitalisierung, Künstliche Intelligenz, Kritik, Ethik, Macht und Herrschaft, (Feministische) Wissenschafts- und Techniksoziologie

Keywords, engl.: Digitalization, Artificial Intelligence, Critique, Ethics, Power, (Feminist) Science and Technology Studies



Bianca Prietl holds a PhD in sociology, and currently works at the department for Sociology with a Focus on Innovation and Digitalization at Johannes Kepler University Linz (Austria). Her main areas of expertise are (Feminist) Science and Technology Studies, Gender Studies, Sociology of Knowledge, and Qualitative Social Research. **E-Mail: bianca.prietl@jku.at**

AI ethics und die mannigfaltigen Verwerfungen von Digitalisierung, Datafizierung und KI

In den letzten Jahren mehrten sich Analysen zu den sozialen, politischen und ökonomischen Verwerfungen im Gefolge von Digitalisierung und Datafizierung, insbesondere des Einsatzes Künstlicher Intelligenz (KI): Diese reichen von Fällen algorithmischer Diskriminierung und der Frage, ob Technik sexistisch oder rassistisch sein kann (Noble 2018; Gebru 2019; Prietl 2019a), über sogenannte Filterblasen und die Sorge, dass *social bots* als menschliche Mitdiskutant*innen verkannt werden und Meinungsbildungsprozesse prägen (Pariser 2011; Wooley 2016; Pörsksen 2018; Dutton et al. 2019), bis hin zu Privatheitsverletzungen durch Internet- und Datenkonzerne sowie staatliche Organisationen, die wiederholt Anlass zur Debatte geben, wie diese reguliert werden könnten (Lyon 2004; Leighton et al. 2017; Véliz 2021). All dies hat erhebliche Zweifel an den Emanzipations-, Demokratisierungs-, Dezentralisierungs- und Objektivitätsversprechen digitaler Datentechnologien**[1]** aufkommen lassen (Morozov 2013; Dickel/Schrape 2015; Prietl 2019b). In Reaktion auf das verbreitete Unbehagen, das mit diesen Entwicklungen verbunden ist, ertönt in Politik, Wissenschaft und Wirtschaft derzeit vor allem ein *Ruf nach Ethik*. Während sich eine Ethik der KI (*AI ethics*), auch Digital-, Computer- oder IT-Ethik genannt, gerade erst konstituiert (Dignum 2018), wie jüngst eingerichtete Professuren und Forschungszentren an der Schnittstelle von KI, Digitalisierung und Ethik (so etwa das 2019 an der TU München gegründete *Institut für Ethik in der KI***[2]**) ebenso demonstrieren wie rezente wissenschaftliche Publikationen (etwa das von Dubber und anderen 2020 herausgegebene *Oxford Handbook of Ethics of AI*), nimmt die Hoffnung auf eine moralphilosophische Einhegung der mannigfaltigen Verwerfungen im Gefolge der Digitalisierung derweilen vor allem die Gestalt von Ethik-Ausschüssen und -Gütesiegeln sowie Ethikrichtlinien und -standards an (bspw. das *Gütesiegel des KI Bundesverband e.V.***[3]** oder die *Ethics Guidelines for Trustworthy Artificial Intelligence* der Kommission der Europäischen Union**[4]**, die beide 2019 erlassen wurden). Diese beanspruchen, ‚Regeln' für die Entwicklung digitaler Datentechnologien bereit zu stellen, deren Befolgung ethisch unbedenkliche technische Artefakte garantieren soll. Das *AI Ethics Guidelines Global Inventory***[5]** der deutschen Watchdog-Organisation *Algorithm Watch* dokumentiert deren Aufstieg und zählte im Sommer 2020 bereits 160 Einträge, wovon der größte Teil seit 2018 veröffentlicht wurde. Es ist diese aktuell dominierende Form, in der gesellschaftlich Kritik an digitalen Datentechnologien geübt und zugleich ein Umgang mit dieser Kritik gesucht wird, die Gegenstand des vorliegenden Beitrags ist.

Sozialwissenschaftliche Analysen zu *AI ethics* problematisieren vor allem die starke Einflussnahme industrieller Akteur*innen. Beispielhaft hierfür sind das bereits genannte *Institut für Ethik in der KI* an der TU München, das von Facebook kofinanziert wird, oder die von der EU-Kommission herausgegebenen Ethikrichtlinien, bei deren Erarbeitung Industrievertreter*innen eine tragende Rolle spielten (Nosthoff/Maschewski 2019). Vor diesem Hintergrund bezeichnen kritische Stimmen *AI ethics* auch als großes Ablenkungsmanöver, mit dessen Hilfe die Internet- und Datenindustrie mit ihrem ‚business as usual' fortführe. Statt auf eine Entwicklung in Richtung mehr

[1] Digitale Datentechnologien bezeichnen technische Artefakte, die mit digitalen Daten operieren, beispielsweise KI-Technologien aber auch Mailprogramme oder Tracking-Apps. Sie bilden den Kern aktueller soziotechnischer Transformationsprozesse, die unter dem Stichwort Digitalisierung verhandelt werden (Houben/Prietl 2018).

[2] Siehe online unter: https://www.tum.de/nc/die-tum/aktuelles/pressemitteilungen/details/35188/ (zuletzt: 6. Juni 2021).

[3] Siehe online unter: https://ki-verband.de/wp-content/uploads/2019/02/KIBV_Guetesiegel.pdf (zuletzt: 6. Juni 2021).

[4] Siehe online unter: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (zuletzt: 6. Juni 2021).

[5] Siehe online unter: https://inventory.algorithmwatch.org/ (zuletzt: 6. Juni 2021).

sozialer Gerechtigkeit hinzuwirken, diene die wachsende Zahl an Ethikrichtlinien und -standards vor allem dazu, Wettbewerbsvorteile zu erzielen und politische Entscheidungsträger*innen sowie die Öffentlichkeit davon zu überzeugen, dass rechtlich unverbindliche Selbstverpflichtungen genügten, um den immer augenscheinlicher zu Tage tretenden Verwerfungen von Digitalisierung und Datafizierung zu begegnen. KI-Ethik – so das Kernargument ihrer Kritiker*innen – werde instrumentalisiert, um als zahnlose Kommunikationsstrategie ernstzunehmenden politischen Diskussionen und rechtlichen Regulierungen einen Riegel vorzuschieben (Sloane 2019; Benkler 2019; Rességuier/Rodriguez 2020). Während der vorliegende Beitrag die Skepsis gegenüber einem solchen ‚ethics washing' durchaus teilt, möchte er einen weit seltener betrachteten Aspekt zur Diskussion stellen – nämlich die *Prämissen und konzeptionellen Grundlagen* dieser an Begriffe der Ethik anschließenden Kritik wie Lösungssuche. Es geht ihm darum zu fragen, welche Formen der Thematisierung und Problematisierung von Digitaltechnik in aktuell diskursprägenden *AI ethics*-Initiativen angelegt sind beziehungsweise was sich in deren Kontext über digitale Datentechnologien überhaupt wie sagen, denken und kritisieren lässt – und was nicht.

Die unter dem Label *AI ethics* firmierenden Richtlinien und Standards bilden dabei einen instruktiven Untersuchungsgegenstand, machen sie doch die Werte und Normen, Vorstellungen und Ideen explizit, die digitalen Datentechnologien gegenwärtig zugrunde gelegt werden.**[6]** Wenngleich umstritten ist, wie effektiv Ethikrichtlinien und -standards das Denken und Handeln von Technikentwickler*innen beeinflussen (McNamara et al. 2018), gelten sie doch als „powerful instruments for constructing and imposing a shared ethical frame on a contentious conversation" (Greene et al. 2019, 2129). Für das hier verfolgte Interesse werden sie als *diskursive Elemente* konzipiert (Foucault 1978; Paulitz 2005; Prietl 2019b), die prästrukturieren, wie digitale Datentechnologien betrachtet werden, was legitimerweise von diesen verlangt werden kann, und welche Arten und Weisen ihres Designs, Einsatzes und ihrer Nutzung denkbar sind. Als Teil gesellschaftlicher Diskurse rund um Fragen von Digitalisierung und Datafizierung entfalten sie (macht)produktive Effekte, indem sie bestimmte Wege der Entwicklung und des Einsatzes von digitalen Datentechnologien überhaupt erst bereiten, während sie andere als undenk- und -sagbar ausschließen. Für den vorliegenden Beitrag wurden 16 *AI ethics*-Initiativen, die mittels theoretischem Sampling über das AI Ethics Guidelines Global Inventory generiert wurden, einer diskursanalytischen Reflexion zugeführt, die ihrerseits durch rationalitäts- und herrschaftskritische Perspektiven auf das Zusammenspiel von Wissen, Macht und Technik informiert ist. Dieses Sample umfasst KI-Ethikrichtlinien und -standards, die entlang folgender drei Achsen gestreut sind: (a) Autor*innen beziehungsweise veröffentlichende Organisation (Privatwirtschaft, staatliche Organisation, Wissenschaft, Zivilgesellschaft); (b) geopolitische Reichweite (national, international, global); und (c) Verbindlichkeitsgrad (bindende Vereinbarung, Selbstverpflichtung, Empfehlung).**[7]**

Auf den verbleibenden Seiten wird nun unter Hinzuziehung einschlägiger Literatur die These entfaltet, dass die konzeptionellen Prämissen von KI-Ethikrichtlinien und -standards, wie sie aktuell die gesellschaftliche Digitalisierungskritik prägen, zumindest drei Limitationen aufweisen – insbesonde-

[6] Dabei kann es Diskrepanzen zwischen *talk*, *action* und *decision* (Brunsson 1993) geben, wenn Organisationen konfligierende Anforderungen wie Profitmaximierung und *social responsibility* navigieren (für Verhandlungen von Ethik-Fragen in Unternehmen des Silicon Valley siehe Metcal et al. 2019).

[7] Selbstredend kann damit *nicht* beansprucht werden, Aussagen über *die* Ethik oder auch nur *die KI*-Ethik zu treffen, da beide gleichermaßen umfangreiche wie heterogene Forschungs- und Aktionsfelder darstellen. Stattdessen macht der vorliegende Beitrag die Prämissen der gegenwärtig dominierenden Form, in der Digitalisierungskritik geübt wird beziehungsweise in der nach Lösungen für deren Verwerfungen gesucht wird, zum Gegenstand einer kritischen Reflexion.

re wenn es darum geht, soziale Macht- und Herrschaftsverhältnisse und die Rolle, die digitale Datentechnologien an deren Aufrechterhaltung spielen, zu verstehen und zu hinterfragen. Diese zeigen sich weniger explizit, sondern lassen sich vielmehr in diskursiven Leerstellen finden, also in dem, was *nicht* thematisiert wird: (1) ein *asozialer Handlungsbegriff*, der eine systematische Einbettung von Handeln in soziale Strukturen, Handlungskontexte und symbolische Ordnungen vermissen lässt; (2) ein *individualistisches Problemverständnis*, das die soziale Strukturierung von Technik nicht konsequent reflektiert; sowie (3) eine Fokussierung auf *Fairness als normativer Fluchtpunkt*, der die unterschiedliche Positionierung von Menschen in gesellschaftlichen Herrschaftsverhältnissen nicht ausreichend berücksichtigt.

Limitationen gesellschaftlich dominierender Formen der Digitalisierungskritik und Lösungssuche

1. Handlungsbegriff

Im Kontext der Technik- und Maschinenethik gibt es schon länger Bemühungen, den Kreis moralisch verantwortlicher Entitäten theoretisch-begrifflich auf Nicht-Menschen, genauer Maschinen und insbesondere KI, auszuweiten (u.a. Adam 2008). Denn wo die Absicht beziehungsweise das Handlungsziel den zentralen Referenzpunkt für die moralische Beurteilung einer Handlung bildet und die (moralische) Verantwortung für ‚gutes' oder ‚schlechtes' Handeln bei den einzelnen Handlungsträger*innen verortet wird, die als dessen Urheber*innen konzipiert werden, stößt der in der westlich-eurozentrischen Geistesgeschichte etablierte Handlungsbegriff an seine Grenzen (Zwitter 2014, 1f.). Insofern nämlich technischen Artefakten in der Regel nicht die mentalen Voraussetzungen attestiert werden, *absichtsvoll* zu handeln, können sie per definitionem auch nicht moralisch (verwerflich) agieren. Unter dem Label *ethics by design* gibt es zudem Bemühungen, Maschinen mit moralischer Urteilskraft auszustatten, also sogenannte *moral machines* zu entwerfen (Allen et al. 2006; Etzioni/Etzioni 2017; Cervantes et al. 2019). Die Herausforderung liegt dabei nicht nur darin, moralische Urteilsfähigkeit derart zu modellieren, dass diese programmier- und damit maschinell prozessierbar ist, sondern überhaupt zu definieren, was, wann und wie moralisch ‚gutes' Handeln auszeichnet. Im Vordergrund steht damit die Frage nach möglichst universalen Handlungsnormen. Im Kern wird so an einem Handlungsbegriff festgehalten, der ein autonom (und rational) handelndes Subjekt als Träger*in (un)moralischer Handlungen voraussetzt; gleichzeitig werden die sozialen Strukturen, Handlungskontexte und symbolische Ordnungen, innerhalb derer sich dieses Handeln vollzieht, ungleich weniger reflektiert. Dies zeigt sich auch bei den betrachteten Ethikrichtlinien und -standards: Kaum einmal erfolgt hier eine systematische Einbettung von – menschlichem oder maschinellem – Handeln in die sozialen Kontexte, innerhalb derer sich dieses entfaltet; vielmehr wird beständig an die Figur des*r autonomen Handlungsträger*in als Adressat*in der postulierten KI-Ethiknormen und -regeln appelliert.

Damit verbunden zeigen sich zumindest drei ‚blinde Flecken' in der gesellschaftlichen Digitalisierungskritik: Die Fokussierung auf vorgeblich willentliche und absichtsvolle Handlungen (einzelner Individuen) vernachläs-

sigt erstens, dass sich Handlungen immer innerhalb und vor dem Hintergrund von zutiefst hierarchischen sozialen Strukturen und symbolischen Ordnungssystemen vollziehen, die den Einzelnen vorgängig und kaum verfügbar sind. Zweitens finden die gleichermaßen unhintergehbaren Konstellationen der Interdependenz zwischen Akteur*innen kaum Berücksichtigung, wobei einzelne Handlungen beziehunsgweise Handlungsträger*innen tendenziell isoliert betrachtet werden. Drittens eignet sich die Figur des autonom-rationalen Handlungssubjekts wenig, um der Vielzahl an Handlungen gerecht zu werden, die sich präreflexiv und auf Basis inkorporierter Deutungs-, Wahrnehmungs- und Handlungsschemata vollziehen, die ihrerseits wiederum an die strukturell-symbolische Gesellschaftsordnung rückgebunden sind.

Demgegenüber fokussierte ein dezidiert *sozialer* Handlungsbegriff darauf, dass handelnde Personen stets gesellschaftlich situiert sind und dass mit den hierarchisch strukturierten Positionen, die sie in der Gesellschaftsordnung einnehmen beziehunsgweise die ihnen zugewiesen werden, bestimmte Normen und Erwartungen ebenso verknüpft sind wie bestimmte Handlungsoptionen, -ressourcen und -zwänge, über die diese nicht abschließend verfügen können – einschließlich dessen, was in einer konkreten Situation eine ‚gute' Handlung darstellt (klassisch: Weber 2008, 3f.; Bourdieu 1987, 97ff.; Emirbayer/Mische 1998). Vor diesem Hintergrund scheint eine analytische Verschiebung in der gesellschaftlichen Debatte um Digitalisierung notwendig – das heißt eine Fokussierung weniger auf die einzelnen Personen, auf technische Artefakte und ihr ‚Tun' als vielmehr auf die diese prästrukturierenden sozialen Instanzen, allen voran auf soziale Strukturen, gesellschaftliche Institutionen und kulturelle Ordnungen inklusive ihrer Macht- und Herrschaftsgefüge. Konkret hieße das etwa, Fälle algorithmischer Diskriminierung nicht primär als Ergebnis moralisch verwerflicher Handlungen Einzelner oder bedauernswerte Einzelfälle technischen Versagens zu betrachten, sondern sie konsequent als Effekt und zugleich selbst Phänomen gesellschaftlicher Macht- und Herrschaftsverhältnisse in den Blick zu nehmen. Beispielsweise wäre dann zuallererst (an)zuerkennen, dass ein Großteil der Bilddatenbanken, die im Kontext maschinellen Lernens zum Einsatz kommen, Personen aus dem Globalen Norden überrepräsentiert, unter anderem mit dem Resultat, dass ein anhand dieser Trainingsdatensätze entwickeltes Tool zur automatisierten Identifikation von Hautkrebs eine höhere Treffsicherheit bei Menschen mit hellerer Haut aufweist (Zou/Schiebinger 2018, 325). Entsprechend wären es auch diese historisch etablierten und strukturell wie symbolisch verankerten Asymmetrien in den gesellschaftlichen Technik- und Un/Sichtbarkeitsverhältnissen, die als zugrundliegende *soziale* Phänomene ins Zentrum der Kritik gerückt und bei der Lösungssuche adressiert werden müssten, sollen digitale Datentechnologien bestehende gesellschaftliche Ungleichheitsrelationen nicht unhinterfragt fortschreiben.

2. Problemverständnis

In enger Verbindung damit steht ein *individualistisches Problemverständnis*, das gegenwärtig in der gesellschaftlichen Digitalisierungskritik deutungsmächtig ist. So heben die analysierten Ethikrichtlinien und -standards zuvorderst auf isolierte ‚Fehler' (von Technik oder Mensch) ab, für die es punktuelle und bevorzugterweise technische ‚Lösungen' zu finden

gilt: Prominente Initiativen wie *Discrimination-Aware Data-Mining* oder *Fairness, Accountability and Transparency in Machine Learning* bemühen sich etwa darum, die jeweils als ursächlich für vor allem diskriminierende Algorithmen identifizierten Fehlerquellen auf der Ebene der Technik zu beheben, das heißt durch Entwicklung besserer digitaler Datentechnologien.**[8]** Diese werden so als isolierte Entitäten betrachtet, die ‚nur' von Expert*innen optimiert werden müssten. Unter dem Stichwort *ethics for design* (Bostrom/Yudkowsky 2014; Filipovic et al. 2018) werden außerdem Ethikrichtlinien, -standards sowie Verhaltenskodizes für Technikentwickler*innen und IT-Unternehmen formuliert, deren Einhaltung die Entwicklung ‚guter' Digitaltechnik garantieren sollen. Beide Ansätze zeugen von einer tendenziell technikdeterministischen und -solutionistischen Haltung, insofern sie genuin soziale Probleme als technische re-definieren und „better building" zum einzig legitimen Weg vorwärts erklären (Greene et al. 2019, 2122ff.). Demgegenüber werden gesellschaftlich-politische Auseinandersetzungen darüber, ob und welche Technologien überhaupt wofür und in welchen Kontexten wünschenswert sind, nicht als Option aufgerufen. Einhergehend damit wird in den betrachteten Richtlinien und -standards weder die soziale Strukturierung von Technik noch die strukturierende Rolle von Technik selbst konsequent reflektiert.

Im Unterschied dazu betonte ein Verständnis von (digitaler) Technik als *soziotechnisches* Phänomen, wie es etwa die (feministische) Wissenschafts- und Technikforschung stark macht, die unauflösliche Verflochtenheit von Technik und Gesellschaft, Materialität und Semiotik (u.a. Barad, 2003; Haraway 2004; Weber 2017, 361ff.). Vor diesem Hintergrund werden zumindest zwei Limitationen eines individualistischen Problemverständnisses deutlich: Zum einen hält dieses an der weit verbreiteten Hoffnung fest, dass neutrale und objektive digitale Datentechnologien möglich seien, wenn denn erst alle Fehler behoben sind. Damit fallen aktuelle Debatten um Digitalisierung und KI immer wieder hinter die zentrale Einsicht der Science and Technology Studies zurück, wonach Technik stets ‚politisch' ist (Winner 1980), nämlich in soziale Macht- und Herrschaftsverhältnisse eingebettet, diese materialisierend und reproduzierend. Würde die Idee einer neutralen Erkenntnisposition hingegen aufgegeben (Haraway 1988), würden Fragen danach virulent, wer an der Entwicklung von Digitaltechnik (nicht) beteiligt ist, wessen Ideen, Wünsche und Bedarfe bei ihrer Gestaltung (nicht) berücksichtigt werden, wie deren epistemologisch-ontologischen Grundlagen aussehen und verobjektiviert werden (Suchman 2008; Weber/Prietl 2021). Zudem sensibilisierten neomaterialistische und postsoziale Perspektiven, die auch nicht-menschlichen Entitäten *agency* attestieren, dafür, dass Handlungsmacht als stets kontingentes Ergebnis des Zusammenwirkens einer Vielzahl von interdependenten – je nach Theorieperspektive auch: intra-dependenten – menschlichen wie nicht-menschlichen Entitäten verstanden werden muss (Barad 2003; Haraway 2004). Damit wird nicht nur die Vorstellung vom Menschen als alleinig handlungsfähig abgelehnt, sondern auch die Idee einer verteilten *agency* geprägt, die isolierte Fehler- und Lösungsbetrachtungen als hochgradig unterkomplex erscheinen lässt (auch: Amoore 2020).

Für die gesellschaftliche Digitalisierungskritik folgte daraus, digitale Datentechnologien nicht ‚nur' als etwas zu betrachten, das Entscheidungen informiert oder automatisiert trifft, sondern (an)zuerkennen, dass diese Tech-

[8] Viele Elemente, die in KI-Ethikrichtlinien Eingang finden, lassen sich vergleichsweise einfach mathematisch operationalisieren und mittels technischer Lösungen adressieren, sodass einige Unternehmen auch bereits zielgenaue technische *fixes* anbieten (Hagendorf 2020, 103).

nologien ganz grundsätzlich und zentral in die Produktion von Bedeutung und sozialen Ordnungen involviert sind (auch Hoffman 2019). Sie haben nicht nur Anteil daran, welche Positionen Menschen zugewiesen werden, sondern welche Positionen es überhaupt gibt, in welchem (hierarchischen) Verhältnis diese zueinanderstehen und mit welchen Handlungsoptionen sie versehen sind. So berechnet der sogenannte AMAS-Algorithmus des Österreichischen Arbeitsmarktservices die Wahrscheinlichkeit, mit der als arbeitssuchend gemeldete Menschen erfolgreich in den Arbeitsmarkt reintegriert werden, um diese entsprechend einer von drei Gruppen zuzuordnen und differenzierten Zugang zu sozialstaatlichen Unterstützungsleistungen zu gewähren. Dass, welche und wie Menschen dabei als beispielsweise ‚schlecht vermittelbar' eingestuft und entsprechend von bestimmten Ressourcen abgeschnitten werden, ist dabei das Ergebnis des kontextspezifischen Zusammentreffens von unter anderem einem bestimmten politischen Ziel (Optimierung des Ressourceneinsatzes im Kontext neoliberaler Wohlfahrtsstaatreformen), einem statistischen Model, das die Grenzen zwischen den einzelnen Gruppen so zieht, dass seine Gesamttrefferquote optimiert wird, sowie einem Arbeitsmarkt, auf dem es in der Vergangenheit Mütter mit Kindern ebenso schwerer hatten, einen Job zu finden, wie Ältere (für Details: Allhutter et al. 2020). Es sind entsprechend all diese – und weitere – Elemente, die *zusammen* und in ihrem Zusammenspiel betrachtet und diskutiert werden müssten, soll das ‚Problem' der von diesem Algorithmus vorgenommenen Diskriminierungen grundlegend verstanden und systematisch adressiert werden. Damit geht es um mehr und Grundsätzlicheres, als in den vielzähligen Diskussionen um algorithmische *biases* oft suggeriert – nämlich nicht ‚bloß' darum, maschinelle Lernalgorithmen mit ‚besseren' (Trainings-)Datensätzen ‚zu füttern', sondern deren alles andere als triviale Verflochtenheit mit sozio-kulturellen, politisch-ökonomischen und strukturell-materiellen Bedingungen in den Blick zu nehmen.

3. Normativer Fluchtpunkt

In der Vergangenheit hat gerade die unterschiedliche Beurteilung von Personen aufgrund von rechtlich geschützten Kategorien wie Religion, Alter, Geschlecht oder sexuelle Orientierung durch digitale Datentechnologien für öffentliche Aufregung gesorgt. Dennoch sind Ungleichheit, Macht und Herrschaft keine zentralen Themen in den betrachteten Ethikrichtlinien und -standards; im Vordergrund stehen hingegen Transparenz und Nachvollziehbarkeit, (Daten-)Sicherheit, Zurechenbarkeit, Verlässlichkeit und Vertrauenswürdigkeit (auch: Daly et al. 2019; Greene et al. 2019; Hagendorff 2020). Wo die Reduzierung von Ungleichheit doch als Ziel ‚guter' Digitaltechnik ausgelobt wird, wird dieses allerdings nur selten näher operationalisiert – und wenn, dann in *Fairness* übersetzt. Dabei wird zumeist die Gleichbehandlung von Menschen zum normativen Fluchtpunkt erklärt und Nicht-Unterscheidung zum Garanten für Gerechtigkeit erhoben.

Die beobachtbare Rahmung von Gleichheit als Fairness birgt erneut zwei Verkürzungen: Einerseits wird – ähnlich wie in vergleichbaren Anti-Diskriminierungsbemühungen (Hoffman 2019, 905ff.) – eine isolierte Betrachtung von Differenzierungskategorien nahgelegt, also die Problematisierung von sexistischer, rassistischer *oder* Altersdiskriminierung. Aus dem Blick geraten damit die unter dem Stichwort Intersektionalität thematisierten emer-

genten Effekte des Zusammenspiels verschiedenerer Ungleichheitsrelationen – so etwa dass Bilderkennungstechnologien Schwarze Frauen deutlich seltener akkurat erfassen können als ‚weiße' Frauen, aber auch als Schwarze Männer (Zou/Schiebinger 2018, 325). Zudem bedeutet Gleichbehandlung im Sinne von Nicht-Unterscheidung da, wo Menschen mehr oder weniger privilegierte Positionen in der Gesellschaft innehalten, realiter eine Nicht-Berücksichtigung der mit diesen Positionierungen verbundenen unterschiedlichen Ausgangslagen. Der im US-amerikanischen Strafvollzug eingesetzte COMPAS-Algorithmus ist etwa dafür in die Kritik geraten, dass er Afroamerikaner*innen eine höhere Rückfallwahrscheinlichkeit attestiert als ‚weißen' Angeklagten – ohne die Kategorie ‚race' überhaupt explizit einzukalkulieren. Erklärt wird dies unter anderem dadurch, dass die Treffsicherheit des Algorithmus bei ‚Weißen' bedeutend höher ist als bei Schwarzen, dass Schwarze also öfter als *false positives* ausgewiesen werden (Angwin et al. 2016). Obwohl nun im Sinne der Fairness daran gearbeitet wird, den Algorithmus für beide Gruppen gleichermaßen treffsicher zu machen (z.B. Corbett-Davies/Goel 2018), ist dennoch davon auszugehen, dass Afroamerikaner*innen weiterhin öfter als rückfallgefährdet eingestuft werden – und zwar weil viele der hierfür als ursächlich modellierten Faktoren wie Arbeitslosigkeit oder niedrige Bildung *nicht* gleichmäßig auf die Bevölkerung verteilt sind; vielmehr sind Afroamerikaner*innen hier aufgrund rassistischer sozialer Strukturen und kultureller Ordnungen überrepräsentiert. Werden letztere nicht berücksichtigt, werden diese nicht nur unsichtbar gemacht und hinter vorgeblich neutralen Berechnungsmodellen ‚versteckt', sondern vor allem unhinterfragt fortgeschrieben.

Anstatt auf Gleichbehandlung und die Idee einer neutralen Technik zu setzen, wäre in der gesellschaftlichen Debatte um Digitalisierung deshalb stärker darauf zu insistieren, Digitalisierung dezidiert mit einem politischen Anliegen zu verknüpfen und in der Technikentwicklung gezielt auf den Abbau von sozialen Hierarchien und gesellschaftlichen Machtasymmetrien hinzuwirken (Paulitz/Prietl 2019, 13; D'Ignazio/Klein 2020).

It is about power, stupid!

Der vorliegende Beitrag hat in einer diskursanalytischen Reflexion skizziert, wie in der aktuell durch Ethikrichtlinien und -standards dominierten gesellschaftlichen Digitalisierungskritik weitestgehend unhinterfragt an individualistische moralphilosophische Prämissen angeschlossen wird, wie sie die westlich-eurozentrische Geistesgeschichte prägen (auch: Jaume-Palasi 2019, 483). Damit gehen erhebliche Limitationen einher, wenn es darum geht, die Bedeutung, die digitale Datentechnologien für die Aufrechterhaltung gesellschaftlicher Macht- und Herrschaftsverhältnisse haben, systematisch in den Blick zu nehmen, für eine gesellschaftlich-politisch Diskussion zu öffnen und nach Möglichkeiten ihres Abbaus zu suchen. Anstatt nämlich die sozialen Strukturen und kulturellen Ordnungen zu thematisieren, innerhalb derer die Verwerfungen von Digitalisierung und Datafizierung überhaupt erst auftreten, die diese prästrukturieren, privilegieren, legitimieren und auch jenseits von ‚verbesserten' und ‚gut' handelnden Akteur*innen fortführen, lenken diskursprägende KI-Ethikrichtlinien und -standards die

gesellschaftliche Aufmerksamkeit zuvorderst auf einzelne sprichwörtlich ‚schwarze Schafe', die als ursächlich verantwortlich für die Verwerfungen digitaler Datentechnologien identifiziert werden, und setzen auf deren ‚Korrektur' als ‚Lösung'.

Vor diesem Hintergrund scheint die beobachtbare Hinwendung zu (KI-) Ethikrichtlinien weder ein ‚Allheilmittel' gegen die mannigfaltigen Verwerfungen im Gefolge von Digitalisierung und Datafizierung noch ein neutrales Unterfangen. Nach Foucault (1978) entfaltet sich die Macht in und durch Wissen, sie operiert dezentral und vielgestaltig, und verfestigt sich erst im Vollzug zu übergeordneten Herrschaftsstrukturen. Im Anschluss hieran lassen sich (KI-)Ethikrichtlinien und -standards als im Kern macht- und herrschafts*konservierend* beschreiben, insofern die hier diskursprägenden Prämissen die sozialen Macht- und Herrschaftsverhältnisse, innerhalb derer digitale Datentechnologien situiert sind und denen sie dienen, weitestgehend unangetastet lassen – nämlich die von einigen wenigen, vornehmlich privatwirtschaftlichen aber auch staatlichen Organisationen, aufgrund der extremen Ressourcenintensität und hohen Skaleneffekte der heutzutage tonangebenden datenbasierten KI, etablierte, monopolartige Vorherrschaft (Srnicek 2018). Diese Konstellation erlaubt es den wenigen dominierenden Akteur*innen, digitale Datentechnologien primär zur Verfolgung eigener Interessen zu entwickeln und einzusetzen, konkret: „profit (for a few), surveillance (of the minoritized), and efficiency (amidst scarcity)" (D'Iganzio/Klein 2020, 41).

Deshalb soll abschließend für eine Stärkung rationalitäts-, macht- und herrschaftskritischer Perspektiven in den gesellschaftlichen Verhandlungen von Digitalisierung und KI plädiert werden. Daraus folgte, grundsätzlich und möglichst breit darüber zu diskutieren, wer an der Entwicklung digitaler Datentechnologien wie beteiligt ist, zu welchen Zielen und Zwecken diese entworfen werden, auf welche Interessen in ihrer Gestaltung eingegangen wird, und wo, wie und zu wessen Vor- beziehungsweise Nachteil diese eingesetzt werden.

Literatur

Adam, A. (2008) Ethics for things. In: *Ethics and Information Technology* 10: 149-154.

Allen, C.; Wallach, W.; Smit, I. (2006) Why Machine Ethics? In: *IEEE Intelligent Systems* 1541-1672(6): 12-17.

Allhutter, D.; Cech, F.; Fischer, F.; Grill, G.; Mager, A. (2020) Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective. In: *frontiers in Big Data* 3(5): 1-17.

Amoore, L. (2020) *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*. Durham: Duke University Press.

Angwin, J.; Larson, J.; Surya, M.; Kirchner, L.; Parris, T. Jr. (2016) Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks. In: *ProPublica*. https://www.propublica.org/article/machine-bias-riskassess-ments -in-criminal-sentencing (26/10/2020).

Barad, K. (2003) Posthumanist Performativity. Toward an Understanding of How Matter Comes to Matter. In: *Signs* 28(3): 801-831.

Benkler, Y. (2019) Don't let industry write the rules of AI. In: *Nature* 569(7754): 161.

Bostrom, N.; Yudkowsky, E. (2014) The ethics of artificial intelligence. In: Russel, S; Norvig, P. (eds.) *The Cambridge Handbook of Artificial Intelligence*. Cambridge: The Cambridge University Press.

Bourdieu, P. (1987) *Sozialer Sinn*. Frankfurt a. M.: Suhrkamp.

Brunsson, N. (1993) Ideas and actions: Justification and hypocrisy as alternatives to control. In: *Accounting, Organizations and Society* 18(6): 489-506.

Cervantes, J.-A.; López, S.; Rodriguez, L.-F.; Cervantes, S.; Cervantes, F.; Ramos, F. (2019) Artificial Moral Agents: A Survey of the Current Status. In: *Science and Engineering Ethics* 26(2): 501-532.

Corbett-Davies, S.; Goel, S. (2018*) The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*. https://arxiv.org/abs/1808.00023 (06/06/2021).

Daly, A.; Hagendorff, T.; Hui, L.; Mann, M.; Marda, V.; Wagner, B.; Wang, W.; Witteborn, S. (2019) *Artificial Intelligence Governance and Ethics: Global Perspectives*. The Chinese University of Hong Kong, Faculty of Law: Research Paper No. 2019-15.

Dickel, S.; Schrape, J.-F. (2015) Dezentralisierung, Demokratisierung, Emanzipation. Zur Architektur des digitalen Technikutopismus. In: *Leviathan* 43(3): 442-463.

Dignum, V. (2018) Ethics in artificial intelligence: introduction to the special issue. In: *Ethics and Information Technology* 20: 1-3.

Dubber, M; Pasquale, F.; Das, S. (2020) (eds.) *The Oxford Handbook of Ethics of AI*. Oxford: Oxford University Press.

Dutton, W.H.; Reisdorf, B.C.; Blank, G.; Dubois, E.; Fernandez, L. (2019) The Internet and Access to Information about Politics: Searching through Filter Bubbles, Echo Chambers, and Disinformation. In: Graham, M; Dutton, W.H. (eds.) *Society and the Internet*. Oxford: Oxford University Press.

Emirbayer, M.; Mische, A. (1998) What is Agency? In: *American Journal of Sociology* 103(4): 962-1023.

Etzioni, A.; Etzioni, O. (2017) Incorporating Ethics into Artificial Intelligence. In: *Journal for Ethic*s 21: 403-418.

European Commission (2019) *Ethics guidelines for trustworthy AI*. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (27/10/2020).

Filipovic, A.; Koska, C.; Paganini, C. (2018) *Developing a Professional Ethics for Algorithmists*. Gütersloh: Berteslmann Stiftung.

Foucault, M. (1978) *Dispositive der Macht. Über Sexualität, Wissen und Wahrheit*. Berlin: Merve.

Gebru, T. (2019) Race and Gender. In: Dubber, M; Pasquale, F.; Das, S. (eds.) *The Oxford Handbook on AI Ethics*. https://arxiv.org/abs/1908.06165 (26/10/2020).

Greene, D.; Hoffmann, A.L.; Stark, L. (2019) Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. In: *Proceedings of the 52nd Hawaii International Conference on System Sciences:* 2122-2131.

Hagendorff, T. (2020) The Ethics of AI Ethics: An Evaluation of Guidelines. In: *Minds & Machines* 30: 99-120.

Haraway, D. (1988) Situated Knowledge: The Science Question in Feminism and the Privilege of Partial Perspective. In: *Feminist Studies* 14(3): 575-599.

Haraway, D. (2004) A Manifesto for Cyborgs: Science, Technology, and Socialist Feminism in the 1980s. In: dies.: *The Haraway Reader*. New York u.a.: Routledge. 7-45.

Hoffmann, A.L. (2019) Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. In: *Information, Communication & Society* 22(7): 900-915.

Houben, D.; Prietl, B. (eds.) (2018) *Datengesellschaft. Einsichten in die Datafizierung des Sozialen*. Bielefeld: transcript.

D'Ignazio, C.; Klein, L.F. (2020) *Data Feminism*. Cambridge: The MIT Press.

Jaume-Palasi, L. (2019) Why We Are Failing to Understand the Societal Impact of Artificial Intelligence. In: *Social Research: An International Quarterly* 86(2): 477-498.

Leighton, A.; Benbouzid, B.; Brice, J.; Bygrave, L.A.; Demortain, D.; Griffiths, A.; Lodge, M.; Mennicken, A.; Yeung, K. (2017) *Algorithmic Regulation*. London: LSE Discussion Paper 85.

Lyon, D. (2004) Globalizing Surveillance: Comparative and Sociological Perspectives. In: *International Sociology* 19: 135-149.

McNamara, A.; Smith, J.; Murphy-Hill, E. (2018) Does ACM's code of ethics change ethical decision making in software development? In: *ESEC/FSE 2018: Proceedings of the 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering:* 729-733.

Metcal, J.; Moss, E.; boyd, d. (2019) Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics. In: *Social Research: An International Quaterly* 86(2): 449-476.

Morozov, E. (2013) To *Save Everything, Click Here. Technology, Solutionism and the Urge to Fix Problems that Don't Exist*. New York: Public Affairs.

Noble, S.U. (2018) *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

Nosthoff, A.-V.; Maschewski, F. (2019) Alles nur Fake Ethik. In: *Republik,* 22.05.2019. https://www.republik.ch/2019/05/22/alles-nur-fake-ethik (06/06/2021).

Pariser, E. (2011) *The Filter Bubble. How the New Personalized Web is Changing What We Read and How We Think*. London: Penguin Books.

Paulitz, T. (2005) *Netzsubjektivität/en. Konstruktionen von Vernetzung als Technologien des sozialen Selbst*. Münster: Dampfboot.

Paulitz, T.; Prietl, B. (2019) Feministische Innovationstheorien. In: Blättel-Mink, B.; Schulz-Schaeffer, I.; Windeler, A. (eds) *Handbuch Innovationsforschung*. Wiesbaden: Springer VS.

Pörksen, B. (2018) Filter Clash. Die große Gereiztheit der vernetzten Welt. In: *re:publica 18*. https://www.youtube.com/watch?v=o3ei8qVgTtc (06/06/2021).

Prietl, B. (2019a) Algorithmische Entscheidungssysteme *revisited*: Wie Maschinen gesellschaftliche Herrschaftsverhältnisse reproduzieren können. In: *feministische Studien* 2(2019): 303-319.

Prietl, B. (2019b) Die Versprechen von Big Data im Spiegel feministischer Rationalitätskritik. In: *GENDER* 3(2019): 11-25.

Rességuier, A.; Rodriguez, R. (2020) AI ethics should not remain toothless! A call to bring back the teeth of ethics. In: *Big Data & Society*. In: https://journals.sagepub.com/doi/10.1177/2053951720942541 (06/06/2021).

Sloane, M. (2019) Inequality Is the Name of the Game: Thoughts on the Emerging Field of Technology, Ethics and Social Justice. In: *Proceedings of the Weizenbaum Conference 2019*. In: https://www.ssoar.info/ssoar/handle/document/62583 (06/06/2021)

Srnicek, N. (2018) Platform Monopolies and the Political Economy of AI. In: McDonnell, J. (ed.) *Economics for the Many*. London: Verso. 152-163.

Suchman, L. (2008) Feminist STS and the Sciences of the Artifical. In: Hackett, E.J.; Amsterdamska, O.; Lynch, M.; Wajcman, J. (eds.) *The Handbook of Science and Technology Studies*. Cambridge u.a.: MIT Press.

Véliz, C. (2021) *Privacy is Power. Why and how you should take back control of your data*. London: Penguin Books.

Weber, M. (2008) *Wirtschaft und Gesellschaft. Grundriss der verstehenden Soziologie*. Frankfurt a.M.: Zweitausendeins.

Weber, J. (2017) Einleitung. In: Bauer, S.; Heinemann, T.; Lemke, T. (eds.) *Science and Technolgoy Studies*. Berlin: Suhrkamp.

Weber, J.; Prietl, B. (2021) AI in the Age of Technoscience. On the Rise of Data-Driven AI and its Epistem-Ontological Foundations. In: Elliott, A. (ed.) *The Routledge Social Science Handbook of AI*. New York: Routledge. i.E.

Winner, L. (1980) Do Artifacts Have Politics? In: *Daedalus* 109: 121-136.

Woolley, S. C. (2016) Automating power: Social bot interference in global politics. In: *First Monday* 21(4).

Zou, J.; Schiebinger, L. (2018) Design AI so that it's fair. In: *Nature* 559: 324-326.

Zwitter, A. (2014) Big Data ethics. In: *Big Data & Society* 2014(1): 1-6.