Citation: Wagener, Andreas (2022). Innovation von Medienprodukten: Corporate Digital Responsibility und KI-Bias, *Hofer Beiträge zur digitalen Transformation*, Okt. 2022, https://doi.org/10.57944/1051-131

Corporate Digital Responsibility und KI Bias

Unter „Corporate Digital Responsibility"(CDR) wird allgemein die gesellschaftliche Verantwortung von Unternehmen im Rahmen der Digitalisierung verstanden. In Anlehnung und als Weiterentwicklung des Begriffs der CSR („Corporate Social Responsibility") richtet sich der – relativ junge, aber aktuell rapide an Relevanz gewinnende[1] – Terminus auf die Anwendung „digitaler" Technologien und den Umgang mit den daraus resultierenden Konsequenzen für die Gesellschaft und den Einzelnen.[2] Im Kern geht es dabei um die freiwillige Selbstverpflichtung, verantwortlich mit den digitalen Ressourcen umzugehen, was sowohl grundsätzliche Fragestellungen der Unternehmensethik als auch konkrete betriebswirtschaftliche Handlungsfelder im Tagesgeschäft berührt.[3]

Dabei stehen insbesondere folgende Themenbereiche im Mittelpunkt des Interesses:

1. die Schonung von Ressourcen bei der Erschaffung digitaler Dienste und Produkte (insbes. Energie)
2. Sozialverträglichkeit und Ermöglichung einer „humanen" Arbeitsumgebung beim Einsatz digitaler Technologie
3. die „Demokratisierung der Digitalisierung": Zugangserleichterung durch Kompetenzaufbau beim Einzelnen und durch die Förderung allgemein zugänglicher digitaler Infrastruktur
4. Datensicherheit, Datenschutz und Verhinderung digitalen Machtmissbrauchs aufgrund erlangter „Datenmacht" (gegen „Überwachungskapitalismus"[4], „Nudging", „invasives" Scoring/Profiling, …)

Und:

5. der verantwortungsvolle Umgang mit KI: Transparenz der Entscheidungsbildung durch KI, Vermeidung von „KI-Bias" und von „Diskriminierung" durch KI.

Bei der Verarbeitung von Daten im Rahmen von KI-Anwendungen bzw. bei der Anwendung von Methoden des maschinellen Lernens besteht die Gefahr, dass es aufgrund der Eigenschaften der dabei verwendeten Datenressourcen zu Verzerrungen kommen kann („KI-Bias"). Ein Beispiel hierfür ist der Prozess der Vorqualifizierung von Bewerbungsdokumenten. Dabei hat sich herausgestellt, dass vorhandene Muster in der bisherigen Belegschaft von der KI erkannt und in den Auswahlprozess einbezogen wurden, was dazu führt, dass z.B. ein Unternehmen mit wenig Frauen in Führungspositionen

[1] Aktuell laufen verschiedene Initiativen der Bundesministerien für Umwelt (BUMV): https://www.bmuv.de/service/veranstaltungen/veranstaltung und Justiz (BMJV): https://www.bmj.de/DE/Themen/FokusThemen/CDR_Initiative/CDR_Initiative_node.html

[2] BMJV (2018): CDR-Initiative und Nachhaltigkeitsziele. https://www.bmj.de/DE/Themen/FokusThemen/CDR_Initiative/_downloads/cdr_nachhaltigkeitsziele.pdf

[3] CSR-News (2018): Corporate Digital Responsibility. https://csr-news.org/2018/06/20/corporate-digital-responsibility/

[4] Zuboff, Shoshana (2015). Big other: surveillance capitalism and the prospects of an information civilization. Journal of Information Technology, 30(1), 75-89. https://doi.org/10.1057/jit.2015.5

oder geringem Anteil von Arbeitskräften mit Migrationshintergrund eine entsprechende Personalpolitik auch in die Zukunft festschreiben könnte, wie ein entsprechendes Projekt von Amazon zeigt (welches dann entsprechend auch nicht umgesetzt wurde).[5] Künstliche Neuronale Netze müssen im Wege maschineller Lernverfahren mit großen Mengen an Daten trainiert werden. Aber Daten sind eben zwangsläufig stets vergangenheitsbezogen und enthalten dann mitunter genau jene Vorurteile und mangelnde Diversität, denen mit CDR-Maßnahmen entgegnet werden soll.

Ähnlich verhält es sich mit Verfahren in der Sprachverarbeitung, die unter anderem für den Betrieb von Chatbots und bei der automatisierten Texterstellung durch KI zur Anwendung kommen. Zur Erreichung eines weitreichenden „Sprachverständnisses" müssen die Systeme mit großen Mengen an Sprachdaten trainiert werden kann. Um an diese zu gelangen, bestehen zwei Möglichkeiten, die jeweils eigene Probleme nach sich ziehen: Zum einen werden gemeinfreie Werke genutzt, welche aufgrund ihres Alters - In Deutschland endet die Schutzfrist 70 Jahre nach dem Tod des Urhebers - meist in besonderer Weise überholte Sprachstile und auch nicht selten überkommene gesellschaftliche Vorstellungen und Motive beinhalten. Zum anderen werden allgemein verfügbare, aktuelle Inhalte aus dem Netz genutzt, wie Blogs und Forenbeiträge, welche keinerlei „manueller" Qualitätssicherung unterliegen und damit potenziell auch extreme Positionen beinhalten können. Zwar lassen sich über automatisierte Prozesse bestimmte Inhalte – wie etwa Pornografie - herausfiltern. Diese Verfahren basieren aber meist auf vorab definierten oder identifizierten Sperrbegriffen und sind nicht in der Lage, subtiler gelagerte Probleme bei Duktus und Stilistik zu erfassen. Die „Antiquiertheit" der Quellen führt unter anderem dazu, dass Sprachmodelle, deren Funktion in der Vervollständigung von Sätzen bzw. dem Weiterführen von Dialogen liegen, entsprechend auch heute als überkommen empfundene Geschlechterrollen reproduzieren. So werden Männer etwa standardmäßig mit Berufen wie Manager, Arzt oder Programmierer in Verbindung gebracht, Frauen hingegen eher als Hausfrau, Sekretärin oder Krankenschwester klassifiziert. Auch die Hautfarbe führt regelmäßig zu einer Vorprägung von KI, zum Beispiel hinsichtlich des Berufes, den eine Person womöglich ausüben oder auch welchen sozialen Status sie inne haben könnte.[6]

Im Sinne eines wirkungsvollen CDR-Konzeptes müssen Unternehmen, die auf Verfahren des maschinellen Lernens zurückgreifen oder entsprechende Produkte einsetzen, sicherstellen, dass diese Verwerfungen eliminiert werden. Allerdings sind damit in der Praxis erhebliche Herausforderungen verbunden: Technisch gibt es bereits erste Ansätze, KI-Bias auch mit KI-Methoden zu bekämpfen, indem man beispielsweise – händisch, über Supervised Learning – entsprechende Texte identifiziert und als „unerwünschte" Ergebnisse etikettiert. In einem iterativen Prozess, immer begleitet durch die menschliche Überprüfung, könnte man damit dann eine Trainingsgrundlage für die KI schaffen, die Sprachdaten und Texte auf ihren Bias-Gehalt überprüft und entsprechend anpasst. Die Forschungspraxis steht in diesem Bereich allerdings noch ganz am Anfang. Und gerade die enge Verwebung von Mensch und maschinellem Lernen birgt natürlich neue Bias-Problematiken.

Grundsätzlich erweist sich bereits die Identifizierung anwendbarer Wertmaßstäbe als schwierig: Nicht immer erscheint die Sachlage so eindeutig wie in den zuvor skizzierten Beispielen: Was allgemein akzeptabel oder wünschenswert erscheint und welche Entwicklungen zu kritisieren sind, ist

[5] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G
[6] https://www.ethikrat.org/fileadmin/PDF-Dateien/Veranstaltungen/anhoerung_25.02.2021_Luxburg.pdf

schließlich Gegenstand permanenter „öffentlicher“ Verhandlung und einem steten Wandel unterworfen. Die viel beklagte gesellschaftliche Fragmentierung erschwert die Extraktion eines allgemein gültigen Wertmaßstabes zusätzlich. Politische Diskurse bilden sich damit automatisch auch in der konkreten Anwendung von KI ab.

Die Berücksichtigung von CDR in diesem Kontext erfordert die Entwicklung eines entsprechenden Instrumentariums, das sowohl die technischen Belange als auch die organisationalen und betriebswirtschaftlichen Anforderungen bedient. Eingebettet in ein zu ermittelndes ethisches Grundgerüst gilt es die Voraussetzungen für ein zielgerichtetes CDR-Management zu schaffen. Dabei müssen die Grenzen zwischen (gewinnorientiertem) Management und technischer Umsetzung sowie zwischen Unternehmertum und gesellschaftlichen Anliegen zwangsläufig verwischen. Wie kann es gelingen, Betriebswirtschaftliches mit dem Gesellschaftlichen zu vereinen? Welche – konkreten – technischen Ansätze sind dabei sinnvoll, um CDR exemplarisch im Kontext von KI-Bias und der Sprachverarbeitung sicherzustellen? Welche strukturellen Maßnahmen gilt es darüber hinaus zu ergreifen? Und wie kann dies alles in eine übergreifende CDR-Strategie eingebettet werden? All dies sind Fragen, welche die derzeitigen Überlegungen zu CDR flankieren müssen. Es ist eine Sache, allgemeine Ziele innerhalb dieses ohne Zweifel sinnvollen Klassifizierungsrahmens aufzustellen. Aber die praktische Umsetzung stellt Forschung, Unternehmen und Politik vor enorme Herausforderungen.