Projektbericht zu fairer KI

Autor_innen: Susanne Wolf-Eberl,[a] Paul Wolf, [a] Peter Reichl, [b] Michael Funk, [b] Christian Löw [b]

Reviewer_innen: Julian Anslinger, [c] Christopher Frauenberger, [b, d] Anita Thaler [c]

[a] Research & Data Competence
[b] Research Group Cooperative Systems, Faculty of Computer Science, University of Vienna
[c] IFZ – Interdisziplinäres Forschungszentrum für Technik, Arbeit und Kultur
[d] Center for Human-Computer Interaction, University of Salzburg

Gefördert im Rahmen des Ideen Lab 4.0 (2019) der FFG.

Inhalt

1. Einführung

Im vorliegenden Bericht erfolgt eine Zusammenfassung der Projektergebnisse zum Themenbereich „Faire KI“.

1.1. Aufgabenstellung

Um innerhalb des Projekts dAIalog.at Möglichkeiten partizipativer Technikgestaltung in Bezug auf Künstliche Intelligenz (KI) und insbesondere deren Fairness kritisch zu diskutieren, bedarf es zunächst einmal einer ersten Auseinandersetzung mit den Besonderheiten von und Ansprüchen an faire künstliche Intelligenz. Dies beginnt mit entsprechender begrifflicher Arbeit, in diesem Fall insbesondere das Problem einer geeigneten Definition von KI sowie einer Abgrenzung von Fairness (z. B. auch gegenüber einem Begriff der Gerechtigkeit) umfasst. Damit verbunden ist aber auch die Frage, ob und in welchem Maße Fairness ein unbedingt anzustrebender Wert ist oder ob im Gegenteil auch faire KI denkbar wäre, die Wertekonflikte impliziert. Schließlich steht damit auch einerseits die technische wie soziale Komplexität von künstlicher Intelligenz (inklusive Mensch-Maschine-Interaktion) und andererseits der ethische Werterahmen (inklusive des Prinzips der Fairness) im Fokus der Betrachtung.

1.2. Überblick über den Projektbericht

Der vorliegende Bericht beschäftigt sich mit den Ergebnissen eines Expert*innen-Workshops sowie verschiedener projektinterner Diskussionen über faire KI. Hierzu werden in **Kapitel 2** zunächst einmal begriffliche Grundlagen gelegt, einerseits hinsichtlich des Verständnisses von KI (einschließlich einer projektintern formulierten Eingangsdefinition), andererseits zum Konstrukt „Fairness“, bevor grundlegende transdisziplinäre Aspekte reflektiert werden. Der technikethische Rahmen wurde zu Projektbeginn in einem internen Workshop geklärt, dessen Resultate in **Kapitel 3** zusammengefasst werden. **Kapitel 4** ist den Ergebnissen des zentralen Expert*innen-Workshops gewidmet. Nach einer Beschreibung des methodischen und inhaltlichen Vorgehens erfolgt zunächst die Validierung der Eingangsdefinition, bevor verschiedene Zugänge zum Thema „KI“ diskutiert werden. Im weiteren Verlauf wird auf positive Erwartungen an faire KI wie auch relevante negative Berührungspunkte eingegangen und schließlich ein erstes Fazit gezogen. Im **Kapitel 5** werden umsetzungsspezifische Aspekte erörtert, insbesondere aus der Sicht österreichischer Unternehmen, bevor wir kurz auch noch auf die Perspektive der EU eingehen. **Kapitel 6** schließt den Bericht mit einem kurzen Fazit ab.

2. Kontext und Begrifflichkeit

Innerhalb des Projekts dAIalog.at ist dieser Bericht der Etablierung des ethischen Kontextes für das Gesamtprojekt gewidmet. Neben einer grundsätzlichen Reflexion wurde darin die Klärung des Begriffes einer „fairen KI“ insbesondere in Form von internen und externen Diskussionen bzw. Workshops vorangetrieben, die mit ausgewählten Stakeholder*innen durchgeführt wurden.
Der vorliegende Bericht reflektiert dabei insbesondere die Ergebnisse der Task 4.1, in deren Rahmen ein interdisziplinärer begrifflicher und konzeptioneller Rahmen zur Schaffung vertrauenswürdiger KI-Systeme basierend auf ethischen und RRI-Prinzipien zu formulieren war.

2.1. Zur Definition von Künstlicher Intelligenz

Es existiert keine einheitliche und allgemein akzeptierte Definition von „künstlicher Intelligenz“ (KI), zudem wird der Begriff etwa von der Enquete-Kommission des Deutschen Bundestages als „sperrig“ und „emotionsbeladen“ charakterisiert, so dass diese in ihrem Abschlussbericht bewusst auf eine Definition verzichtet und sich mit einer „Begriffsklärung“ zufriedengibt (Deutscher Bundestag 2020, S. 1). Demgegenüber schlägt die hochrangige Expertengruppe der EU folgende Definition vor (zitiert nach COM(2020), S. 19 Fußnote 47):

> *„KI-Systeme sind vom Menschen entwickelte Software- (und möglicherweise auch Hardware-) Systeme, die in Bezug auf ein komplexes Ziel auf physischer oder digitaler Ebene agieren, indem sie ihre Umgebung durch Datenerfassung wahrnehmen, die gesammelten strukturierten oder unstrukturierten Daten interpretieren, Schlussfolgerungen daraus ziehen oder die aus diesen Daten abgeleiteten Informationen verarbeiten und über die geeigneten Maßnahmen zur Erreichung des vorgegebenen Ziels entscheiden. KI-Systeme können entweder symbolische Regeln verwenden oder ein numerisches Modell erlernen und sind auch in der Lage, die Auswirkung ihrer früheren Handlungen auf die Umgebung zu analysieren und ihr Verhalten entsprechend anzupassen.“*

Diese Definition betont also die Tatsache, dass KI-Systeme von Menschen im Hinblick auf eine Zielerreichung entwickelt werden und sich demnach in die Kulturgeschichte der Zweck-Mittel Zusammenhänge materieller Technik einreihen. Es sind somit von Menschen gesetzte, sozial wie historisch durchaus unterschiedliche Ziele, zu deren Erfüllung KI-Systeme beitragen sollen. Die EU-Definition nimmt den Zweck der KI, also zumindest indirekt ihren Mittelcharakter zur Erreichung eines vorgegebenen Ziels, in die Definition auf, während die Eingangsdefinition für den Workshop KI als Mittel zum Zweck und der Problemlösung explizit anspricht. Dabei betont die Definition der EU zugleich die Komplexität des Systems wie auch der damit angestrebten Ziele, ohne dass jedoch hinreichend klar wäre, ab welchem Komplexitätsgrad ein derartiges System seinen Charakter als bloßes Tool verliert. Demgegenüber ist der Österreichische Rat für Robotik und Künstliche Intelligenz (ACRAI) bezüglich einer Definition von KI noch vorsichtiger und geht stattdessen von folgender „begrifflicher Annäherung“ aus:

> *„Künstliche Intelligenz bezeichnet Systeme mit ‘intelligentem’ Verhalten, die ihre Umgebung analysieren und mit einem gewissen Grad autonom handeln.*

> *KI-basierte Systeme sind etwa Sprachassistenten, Bildanalysesoftware, Suchmaschinen sowie Sprach- und Gesichtserkennungssysteme. KI findet sich aber auch in modernen Robotern, autonomen Fahrzeugen und Drohnen wieder. KI nutzen wir täglich, um etwa Texte zu übersetzen, Untertitel in Videos zu erzeugen oder unerwünschte Emails zu blockieren." (https://www.acrai.at/ [30.11.2020])*

Der in diesem Definitionsversuch eingebrachte Begriff des „autonomen Handelns" hilft allerdings nur bedingt weiter, da zunächst zu klären wäre, was Autonomie in diesem Zusammenhang bedeutet. Zudem erscheint die in der ACRAI-Definition vorgenommene Aufzählung möglicher Anwendungen nicht hilfreich.

Daher wurde innerhalb des Projekts zunächst ein eigener Definitionsversuch von KI unternommen, um einen gemeinsamen Rahmen abzustecken und daraus zugleich Motive für einen kritischen Dialog zu gewinnen. Dabei wurde diese Definition bewusst kontrovers formuliert, um innerhalb des interdisziplinären Projektteams und insbesondere auch für den ersten Expert*innen-Workshop einen geeigneten Diskursraum zu öffnen und dort als Folie der kritischen Betrachtungen zu dienen.

Die innerhalb des Projekts erarbeitete Eingangsdefinition von KI umfasst dabei die folgenden drei zentralen Punkte:

- KI bildet Aspekte menschlicher Intelligenz mit Computersystemen nach.
- KI ist ein technisches Mittel zum Zweck und umfasst u.a. rudimentäres Lernen, Selbstkorrektur und Schlussfolgerung.
- KI gibt vor, eigenständig Antworten zu finden und selbstständig Probleme zu lösen.

2.2. Zum Konstrukt „Fairness"

KI-basierte Systeme werfen viele Fragen und Risiken auf, die es einzugrenzen und abzuschätzen gilt. Big Data, maschinelles Lernen und selbstlernende Systeme eröffnen ungeahnte Möglichkeiten im positiven und negativen Sinn. Der Faszination disruptiver Innovationen stehen Ängste vor Kontrollverlust, Ohnmacht und Manipulation gegenüber. Gerade bei einer neuen Technologie wie der KI, die einen tiefgreifenden Paradigmenwechsel bedeutet, kann dabei eine potentielle Bedrohung als besonders gravierend wahrgenommen werden, insbesondere wenn keine aktive Möglichkeit der Einflussnahme oder gar Gestaltung gesehen wird.

Vor diesem Hintergrund ergeben sich sogleich einige zentrale Fragen, wie etwa:

- Was sind die Einflussmöglichkeiten auf eine faire künstliche Intelligenz?
- An welchen Punkten der Entscheidungsketten wird Fairness definiert, ausgehandelt und überprüfbar gemacht?
- Wo und wie kann in Österreich ein „Ökosystem für Vertrauen" (COM(2020)) in KI proaktiv mitgestaltet werden, sodass es zu keiner Bedrohung und letztendlich zu Vertrauensverlust kommt?

Hierfür gilt es zunächst zu klären, für welche KI-Anwendungen Fragen der Fairness überhaupt relevant sind. Hierfür hat Katharina Zweig, die sich aus Sicht der Sozioinformatik mit ethisch relevanten algorithmischen Entscheidungssystemen beschäftigt, ein sog. „Algoskop“ als Werkzeug entwickelt, um diejenige Art von Software herauszufiltern, die besonders auf dem Prüfstand stehen: algorithmische Systeme, die unmittelbar über Menschen entscheiden oder Entscheidungen fällen, die Menschen mittelbar betreffen (Zweig 2019). Darunter fallen ihr zufolge insbesondere all jene Algorithmen, die in signifikantem Ausmaß

- über Menschen,
- über Ressourcen, die Menschen betreffen, oder
- über Ressourcen, die die gesellschaftliche Teilhabemöglichkeit von Personen ändern,

entscheiden.

Hierbei stellt sich allerdings sofort die Frage, inwieweit Algorithmen als Operationalisierung sozialer Konstrukte aufzufassen sind, und wenn ja, wie diese Operationalisierung so gelingen kann, dass sie mit Grundrechten und Menschenrechten vereinbar und fair ist. Was bedeutet es, wenn auf Basis von Algorithmen Schlussfolgerungen aus Daten abgeleitet werden, die falsch sein können und zugleich irreversible Entscheidungen zur Folge haben, die schwer beeinsprucht werden können? Genügt es, den Anspruch an Fairness auf der Ebene von Regulierung und Kontrolle zu stellen, oder muss er bereits bei gesellschaftlichen Zielen und Werten ansetzen, welche die Gestaltung von KI beeinflussen?

Unfairness kann in diesem Zusammenhang mehrere Quellen haben. Zum einen kann sie bereits in der Modellierung eines Problems verankert sein, bei der bereits die grundlegenden ethischen Entscheidungen getroffen werden (vgl. O’Neill 2016). Ein einschlägiges Beispiel hierfür liefert der sogenannte „AMS-Algorithmus“[1] mit seiner impliziten Fortschreibung von Trends aus der Vergangenheit (Allhutter et al. 2020). Zum anderen sind die von einem Algorithmus gelernten Entscheidungsregeln auch von den verwendeten Trainingsdaten und den verwendeten Methoden abhängig. Die Qualität der Trainingsdaten trägt dabei nicht nur dazu bei, dass KI-basierte Systeme fehlerfrei arbeiten, sondern hat auch signifikanten Einfluss auf mögliche Biases und damit Fairness. Werden bestimmte Zielgruppen bevorzugt oder andere vernachlässigt, so kann es zu unzulässigen Diskriminierungen und unfairen Schlussfolgerungen kommen. Daher stellen vollständige und möglichst wenig verrauschte Datenbanken eine wichtige notwendige (aber keineswegs hinreichende!) Voraussetzung für faire KI dar.

Der Anspruch auf Fairness ist insbesondere in einem gesellschaftlichen Aushandlungsprozess zur Abstimmung von fairer KI von Bedeutung. Hier ist zu fragen, worauf dabei zu achten ist und wo sich realisierbare Möglichkeiten ergeben, ethisches Handeln zu

[1] Das österreichische Arbeitsmarktservice will mit seinem „Arbeitsmarktchancen- Assistenz-System“ (AMAS) die Biographien und Arbeitsmarktchancen arbeitssuchender Menschen mittels Algorithmus in einen scheinbar „objektiven” Wert übersetzen, um Empfehlungen abzuleiten. Aufgrund der Kritik, benachteiligte Personen am Arbeitsmarkt zu diskriminieren, landete der Algorithmus vor dem obersten Verwaltungsgericht (siehe: https://netzpolitik.org/2021/oesterreich-jobcenter-algorithmus-landet-vor-hoechstgericht/).

berücksichtigen, um das Vertrauen zu steigern? Hierzu wurden in projektinternen Diskussionen die Aushandlungsprozesse fairer KI thematisiert.

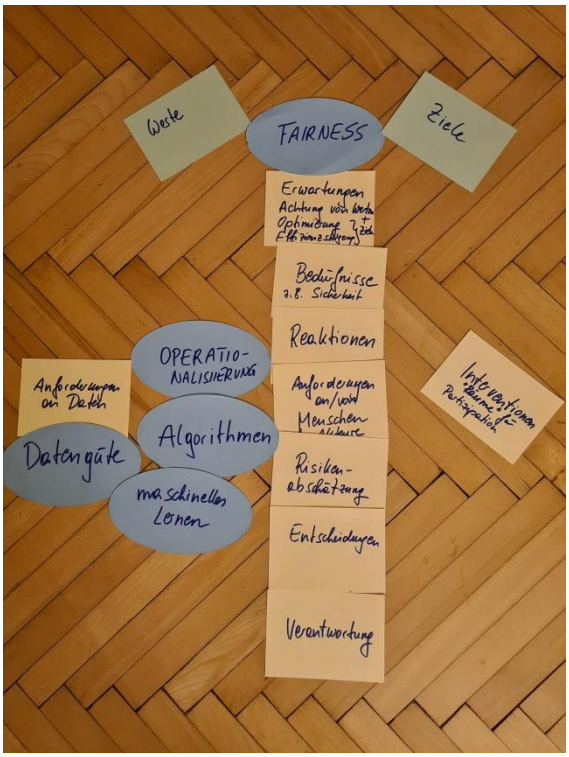

Abbildung 1: Brainstorming zu Aushandlungsprozessen für faire KI

Dabei wurden folgende Dimensionen eines Aushandlungsprozesses identifiziert:

- **Werte** (ethische, moralische, gesellschaftliche, ökologische, soziale, wirtschaftliche ...) und **Ziele** (lebenswerte Zukunft, Nachhaltigkeit, gesellschaftliche Teilhabe ...)
- **Anforderungen an Daten** (Qualitäts- und Gütemaße für Daten und insbesondere Trainingsdaten für Empfehlungssysteme, nachvollziehbare Algorithmen, transparente Operationalisierungskonzepte...)
- **Anforderungen an Menschen** (Fairness durch Transparenz, Nachvollziehbarkeit von Zielsetzungen und Handlungsanweisungen, Akzeptanz von Werten, Gender- und Diversitätsbetrachtung, EU-weite Zertifizierungen, Kennzeichnungen von Black Boxes, Auszeichnungen, Blacklists etc. ...)
- **Bedürfnisse von unterschiedlichen Personengruppen und Akteuren** (Sicherheit, Rechtssicherheit, Komfort, Achtung der Grundrechte, Zufriedenheit, Handlungsfreiheit, Innovation, Profit ...)
- **Reaktionen auf KI-basierte Systeme** (Sichtbarmachen von **positiven Beispielen** und Emotionen wie Faszination und Dankbarkeit für eine Steigerung der Lebensqualität und **Ernstnehmen von Ängsten** und Gefühlen wie Ohnmacht und ausgeliefert sein durch konstruktiven Umgang mit Fehlern ...)
- **Chancen-/Risikenabschätzung** (Problemlösung und Umgang mit Bias, Fehlern, Missbrauch, Manipulation, Diskriminierung, Malsoftware, Betrug, Erpressung ...)

- **Entscheidungen und Verantwortung (**Verantwortungsketten definieren, Handlungsspielräume der unterschiedlichen Akteursebenen samt weißen Flecken ausloten; Kontext des Algorithmus…)
- **Kommunikation und Interaktion** (Offenheit und Verständnis durch Mitgestaltung fördern, Methodenvielfalt ausnutzen und möglichst viele Berührungspunkte mit KI-basierten Systemen ermöglichen, partizipative Technikgestaltung, Diskursräume zwischen Expert*innen und Wutbürger*innen eröffnen, Anlaufstellen für subjektiv empfundene Benachteiligungen ….)

Für jede dieser Dimensionen lassen sich, wie in folgender Tabelle zusammengefasst, nochmals drei Perspektiven unterscheiden: Datenaspekte, Fragen menschlichen Umgangs und sich daraus ergebende zugehörige Interventionsräume.

	Daten	**Mensch**	**Interventionsräume**
		Umgang mit …	
Werte	Erhebung Vertraulichkeit Unterschiedliche Sozialisierungskontexte	Ethischen, moralischen, ökologischen, ökonomischen, gesellschaftlichen, sozialen Vorurteilen Dilemmata	Klärung Operationalisierung Transdisziplinärer Diskurs
Anforderungen	Vollständigkeit, Bias Trainingsdatenbasis Maschinelles Lernen	Güte, Qualitätsmaßen Transparenz von Algorithmen Chancengerechtigkeit/Gender Datenlage, Grundwahrheit	Definition, Verständlichkeit, Schlussfolgerung Überprüfung der Güte Kausalität
Erwartungen und Bedürfnisse	Erkennen von sinnvollen Mustern	Plausibilitätsprüfung, Nachvollziehbarkeit Sicherheit, Teilhabe Vorteil, Benefit, Sinn	Erklärung Schutz Visualisierung
Reaktionen		Faszination Ohnmacht und Ängsten Gleichgültigkeit	positive Verstärkung, Visualisierung Ernstnehmen, Entgegnung Design
Chancen und Risiken	Effizienz, Verlässlichkeit Bias Malware	Problemlösungskompetenz Fehlern und Diskriminierung Risikoabschätzung	Sichtbarmachung Korrektur Orientierungshilfen, "Siegel"
Entscheidung und Verantwortung	Algorithmische Entscheidungssysteme	Handlungsspielräume Verantwortungsketten Regulierungsrahmen	Klärung Einspruchsmöglichkeiten Aushandlungsprozess
Kommunikation und Interaktion		Offenheit Verständnis Vertrauen	Diskurs Wissensvermittlung partizipative Technikgestaltung

Tabelle 1: Dimensionen gesellschaftlicher Aushandlungsprozesse für faire KI

2.3. Gender und Diversity

Künstliche Intelligenz wird im Diversity-Diskurs kritisch thematisiert. Diskutiert wird insbesondere, dass bestehende Ungleichheit und vorhandene Stereotypen Gefahr laufen,

durch Algorithmen weiter verfestigt zu werden, dies aber wenig transparent und schwer nachvollziehbar sei.

Der gesellschaftliche Diskurs über Gendergerechtigkeit ist besonders wichtig in Bezug auf folgende Aspekte:

- Das bestehende Ungleichgewicht der Geschlechter in der Gestaltung von Technik und KI
- Diskriminierung aufgrund KI-basierter unausgewogener Datenbasen und Algorithmen
- Mangelnde intersektionale Betrachtung
- Festschreibung von bestehenden Stereotypen

1.1.1 Bestehendes Ungleichgewicht der Geschlechter in der Gestaltung von Technik und KI

Ein geringerer Anteil an bspw. Frauen im IT-Bereich kann dazu führen, dass gerade in dieser für KI bedeutsamen Entwicklungsperiode der Anteil an Professorinnen, Wissenschaftlerinnen, Gutachterinnen, Kritikerinnen, Rolemodels und Start-Up-Gründerinnen zu langsam wächst. Auf Österreich bezogen bedeutet dies, dass beispielsweise an der TU Wien der Frauenanteil unter Informatik-Absolvent*innen seit Jahren lediglich um die 20% beträgt (s. Abb. 2) und 2019 auf 16% gesunken ist. Der Anteil von Informatik Professorinnen liegt aktuell bei 27%, also 6 Frauen gegenüber 16 Männern (vgl. Genderbericht TU-Wien 2019[2]).

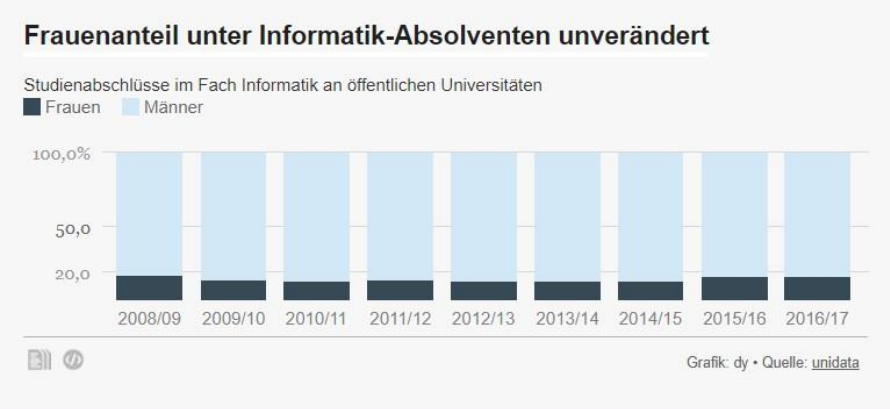

Abbildung 2: Studienabschlüsse im Fach Informatik 2008-2017[3]

Die Teilhabemöglichkeiten an Entscheidungsprozessen auf der Expertinnen-Ebene in IT und Technikbereichen ändert sich daher nur langsam. Viele Entscheidungsgremien scheinen nach wie vor männlich dominiert. So wurde beispielsweise die Gesellschaft für Informatik vom Sachverständigenrat für Verbraucherfragen in Deutschland mit einer

[2] https://www.tuwien.at/fileadmin/Assets/dienstleister/abteilung_genderkompetenz/gender_ressourcen/Zahlen_und_Fakten/Frauenbericht/Frauen_und_Maennerbericht_VII.pdf 2019

[3] Quelle: https://www.derstandard.at/story/2000099099551/warum-so-wenige-frauen-den-code-knacken-wollen

technischen und rechtlichen Betrachtung von algorithmischer Entscheidungsverfahren und der Überprüfung der Machbarkeit für ein Algorithmengesetz im Bereich Verbraucherscoring beauftragt. Es ist anzumerken, dass bei dem Prozess, bei dem es unteranderem um Geschlechterdiskriminierung ging, keine Frau als Gutachterin beteiligt wurde.[4]

Im deutschen Females Founder Monitor[5] wird betont, wie alarmierend es sei, wenn der Anteil an Frauen unter deutschen Gründerinnen und Gründern bei nur 16 % liegt. Hier werden wesentliche Potentiale nicht ausgeschöpft. Eine beträchtliche Hürde ist gemäß dieser Studie unter anderem, dass Frauen weniger Zugang zu Finanzierungsformen bekommen. Wenn diese Tendenz in einem KI-basierten Kreditscoring festgeschrieben wird, setzt dies die Chancen einer gendergerechten Finanzierung weiter herab.

Beispiele wie dieses zeigen, dass es notwendig ist, Gender- und Diversity-Aspekte im Auge zu behalten und Frauen als Expertinnen von anderen Wissenschaftsdisziplinen (Ethik, Recht etc.) sowie als Teil der Zivilgesellschaft, stärker in den gesellschaftlichen Aushandlungsprozess für faire KI einzubinden.

1.1.2 Diskriminierung aufgrund KI-basierter unausgewogener Datenbasen und Algorithmen

Intransparente, fehlerhafte oder inkomplette Datensätze können zu diskriminierenden Ergebnissen führen. Ein Beispiel dafür ist die aktuelle Diskussion über Gesichtserkennungssoftware in Deutschland, die aufgrund fehlender Datenbasis nur weiße Frauen erkannte. Wenn Empfehlungssystemen (bspw. Verbraucherscoring) auf einer männlich dominierten Datenbasis aufbauen, werden andere Geschlechter diskriminiert. In der Datenerfassung nicht ausreichend repräsentierte Gruppen sind gefährdet, durch eine mögliche Verzerrung, ungleich behandelt zu werden (O´Neil 2016)

1.1.3 Mangelnde intersektionale und interdependente Betrachtung

Auch die Verschränkung verschiedener Ungleichheit generierender Strukturdimensionen, wie Geschlecht, Ethnizität, soziales Milieu, Nationalität, Sexualität, Alter, die zu Diskriminierung führen können, spielt im Diskurs eine wichtige Rolle.[6] Eine fehlende oder zu geringe Berücksichtigung von verschränkten, sowie interdependenten Dimensionen kann zu verstärkten Ungleichheitsverhältnissen führen, insbesondere weil mehrfach benachteiligte Personen verstärkt von Diskriminierung bedroht und unzureichend (in Datensätzen) repräsentiert sind.

1.1.4 Festschreibung bestehender Stereotypen

Schließlich können KI-basierte Systeme zu ungleichen Chancen führen, wenn Stereotypen reproduziert und somit bestehende soziale Ungleichheit verstärkt und verfestigt wird. Werden etwa Daten, die Ungleichheit und Ungerechtigkeit abbilden, herangezogen, um KI-Systeme zu trainieren, dann besteht die Möglichkeit, dass beispielsweise rassistische und sexistische Zuschreibungen übernommen werden.

---

[4] https://www.svr-verbraucherfragen.de/wp-content/uploads/GI_Studie_Algorithmenregulierung.pdf

[5] https://femalefoundersmonitor.de/wp-content/uploads/FemaleFoundersMonitor2020.pdf

[6] https://www.politische-medienkompetenz.de/debatte/ki-und-intersektionalitaet/

2.4. Transdisziplinäre Aspekte

Wie aus dem bisher Gesagten unschwer hervorgeht, weist das Thema „Faire KI“ in der hier präsentierten Zuspitzung bereits nicht unerhebliche Komplexität auf. Diese wird im Hinblick darauf besonders eindrücklich, dass jeder der genannten Aspekte mit anderen gesellschaftlichen und technischen Handlungsfeldern in enger Verbindung steht. Außerdem ist die Frage nach Fairness selbst in einem weit über KI hinausweisenden Rahmen menschlich-kultureller Handlungen zu verorten. Faire KI ist breiter als die Summe der hier vorgestellten Teile. Gleiches gilt für Fairness generell, jenseits und diesseits der KI.

Um im hier zugespitzten Rahmen Synergieeffekte zur Formulierung einer problemorientierten Perspektive zu erzeugen, stellte transdisziplinäre Integration (Klein 2008; Pohl et al. 2008) innerhalb des Teams eine besondere Aufgabe dar. Ein Fokus war also nicht nur über das Objekt der Betrachtung, sondern gleichfalls über Kooperationsformen innerhalb der Arbeitsgruppe zu erzielen. In diesem Sinne ging es dann auch tatsächlich um Transdisziplinarität und nicht bloß interdisziplinäres Addieren getrennter Perspektiven (Klein 2017; Mittelstraß 2018). Entscheidend war dabei der normative Zugriff auf das Thema, der über ein bloß beschreibendes Monitoring hinausging.

So gesehen war es von besonderer Wichtigkeit, wie als Nächstes beschrieben, zunächst einmal innerhalb des Projektteams einen technikethischen Rahmen iterativ zu erarbeiten, wobei das gemeinsame Erarbeiten und voneinander Lernen selbst ins Zentrum gerückt wurde.

3. Technikethischer Rahmen

3.1. Vorbemerkungen

Entsprechend der iterativen transdisziplinären Herangehensweise ging es bei der Erarbeitung des technikethischen Rahmens nicht so sehr um das Formulieren eines statischen Wertekatalogs für faire KI. Ein solcher diente eher als „boundary object“ (Bergmann et al. 2010, S. 106-116), also als integrativ wirkender Gegenstand, anhand dessen die Teammitglieder ein gemeinsames Verständnis und eine gemeinsame Sprache finden konnten. Ferner galt es zunächst, im Umgang mit Werten und Ethik disziplinäre Engführungen aufzuheben, wo sie der Problemlösung im Wege standen (Mittelstraß 2018). Aufgrund der kurzen Projektlaufzeit – in der Forschung ist man sich mittlerweile einig, dass wirklich transdisziplinäres Forschen mehrere Rekursionen braucht, die mindestens 3 Jahre gemeinsames Lernen, Sprechen und Arbeiten voraussetzen (Hirsch Hadorn et al. (Hg.) 2008) – wurde daher zu Projektbeginn in einem Workshop zum technikethischen Rahmen einige Iterationen verdichtet und von technikphilosophischer Seite durch Michael Funk moderiert.
Zur Vorbereitung wurde, während der ersten regulären Projektmeetings in lockerer Form Dialoge geführt, bei denen es im Wesentlich noch nicht um Inhalte, sondern um das initiale Aufbauen von Themenfäden ging. Dabei wurde großer Wert auf das Wecken von Nahbarkeit, Neugier und Kreativität gelegt. Inhaltlich wurden dabei bereits (ohne konkreten Ergebnisdruck) im Dialog erste gemeinsame Knoten und teaminterne Synergien, aber auch Reibungspunkte identifiziert. So ist auch der nachfolgende Workshop zur Technikethik nicht als empirische Sozialstudie misszuverstehen, vielmehr war der Moderator stets teilnehmender Dialogpartner und hat bereits vor dem Workshop inhaltlich eingegriffen.

3.2. Interner Workshop zur Technikethik – Format und Ablauf

Der projektinterne Workshop zur Technikethik fand am 2. Juli 2020 statt und von Michael Funk als Fachvertreter der Disziplin Technikethik moderiert. Leider war das ursprünglich vorgesehene Präsenzformat aufgrund der Pandemiebeschränkungen nicht möglich, sodass der Workshop stattdessen online durchgeführt wurde. Am Workshop nahm nahezu das gesamte Projektteam teil, insbesondere waren alle Projektpartner*innen vertreten.
Zur direkten Vorbereitung wurden vor dem Workshop allen Beteiligten drei Fragen gestellt:

1. *“Was wolltest du schon immer einmal über Ethik wissen? “*
2. *“Wer ist dein*e Lieblingsphilosoph*in? Warum? [Anm.: Ich gebe zu, ich selbst mache nicht gerade Freudensprünge wenn mich das jemand fragt. Ich hoffe also inständig Frage 2. ist nicht eure Antwort auf Frage 1.! :-)]“*
3. *“Warum brauchen wir Roboterethik nicht? “*

Diesen Fragen liegt folgende Struktur zugrunde:

- Die erste Frage öffnet den Horizont, soll die Teammitglieder entsprechend ihrer Interessen abholen und einladen, sowie nicht auf direkte Verwertbarkeit im Sinne

der Forschungsfrage zur fairen KI abzielen. Sie dient als hermeneutische Brücke zum Thema.

- Die zweite Frage zielt auf das Vorwissen. Dabei wird durch den Kommentar inkl. Emoji bereits eine dialogisch nahbare Brücke gebaut. Zu keinem Zeitpunkt soll der Eindruck einer isolierten Fragebogenstudie entstehen, sondern das gemeinsame dialogische Lernen sichtbar werden.
- Die letzte Frage zielt nun auf die Verwertbarkeit entsprechend der Forschungsfrage des Projektes ab, jedoch wiederum ohne sie direkt anzusprechen. Ein Perspektivwechsel und damit auch interner Dialog der Beteiligten werden angeregt. Es geht also in der Kürze der drei Fragen um eine dialogisch-hermeneutische Prozessualität, und zwar mit sich selbst und mit den anderen. Dies war im Vorfeld pointiert zu berücksichtigen.

Nachdem jede*r seine Antworten per E-Mail an den Moderator gesendet hatte, wurden diese gesammelt und in das Narrativ eines Fachvortrags zu den ethischen Grundlagen der KI integriert. Hierbei wurde vom Moderator eine besondere integrative Form gewählt, bei der Vortrag samt Folien zwar bereits im Vorfeld angelegt war und daher nicht an den Antworten auf die genannten drei Fragen ausgerichtet werden konnte, sondern umgekehrt jede Antwort ihren Platz im Lauf der Folien fand. So fanden Interessen, Vorwissen und Problemlösungsinteressen aller Beteiligten direkt ihren dialogischen Platz innerhalb des eigentlich bloß fachspezialisiert ausgearbeiteten Referats. So wurde die Rolle der Technikethik/-philosophie als einer disziplinär verankerten und praktisch operationalen Transdisziplin (Funk 2021a) praktisch vorgeführt.
Dies sei beispielhaft anhand der Frage nach der Unterscheidung zwischen zwei Ebenen der KI-Ethik betrachtet, welche in diesem Diskurs eine herausragende Rolle spielt (Funk 2020; Funk 2021b). Zwischen den Folien, die diese Unterscheidung erklärten, wurde (wie in Abbildung 3 ersichtlich) die Antwort eines Teilnehmers platziert.

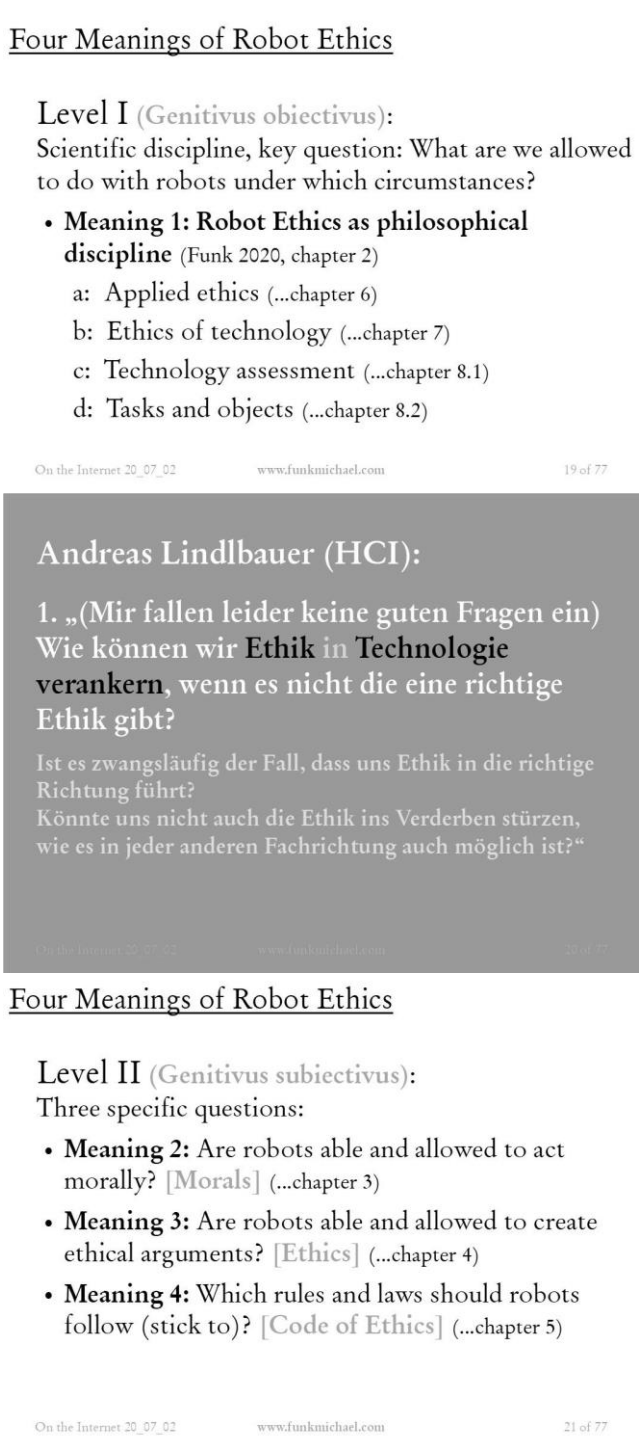

Abbildung 3: Drei Beispielfolien aus dem Workshop-Referat

Die hier in der zweiten Folie dargelegte Frage (als Antwort auf Frage 1) zielt auf Ethik, die nicht von Menschen, sondern von Maschinen ausgeführt wird. Sie veranschaulicht die Analyse der beiden Ebenen aus dem Fachvortrag und führt zugleich in die methodisch relevante Problematik ein, dass es tatsächlich nicht die eine Ethik gibt, sondern mehrere Begründungsformen von deontologischem Ansatz bis hin zum Utilitarismus. Mit der prominenten Platzierung des Beitrags konnte sich nicht nur der betreffende Kollege wiederfinden, sondern auch gleichzeitig seine Unsicherheit – er glaubte ja keine gute Frage zu

formulieren, was in der Tat absolut nicht so ist – abzubauen und ihn in den gemeinsamen Lernprozess zur Problemlösung zu integrieren.
Insgesamt enthielt dieser integrative Ansatz für die Beteiligten diverse Überraschungsmomente, da sie nicht über das Einfügen ihrer Antworten in die Vortragsfolien vorab informiert wurden. Die in den Wochen zuvor stattgefundenen Gespräche und Perspektiven fanden ihren dialogischen Raum und wurden im Anschluss vertieft. Nach einer zweiwöchigen Pause fand eine weitere Nachlese im Gremium statt.
Im Ergebnis bleibt festzuhalten, dass das im Verlauf des Workshops gereifte Verständnis von Ethik als Wissenschaft von der Moral und ihren Unterschieden zu bloß sittlicher Praxis oder formulierten Moralkodizes innerhalb des Projektteams zu diversen Synergieeffekten führte, die sich im technikethischen Rahmen der Analyse fairer KI in Österreich spiegeln. Herausragende Bedeutung erfuhr dabei die wechselseitige Ergänzung normativ-reflexiv arbeitender Philosophie bzw. Geisteswissenschaften und empirischer Sozialforschung – beides durch die Forschungsfrage des Projektes gefordert und iterativ im Workshop behandelt.

3.3. Zusammenfassung und Schlussfolgerungen

Das Projekt dAIalog.at zielt auch auf ein deskriptives Monitoring fairer KI ab. Der zugrundeliegende technikethische Rahmen baut auf einer methodisch-sprachkritischen Anthropozentrik (Funk 2021b; Funk 2021c) auf, um die normativen und empirischen Aspekte der Rede von „Fairness“ aus menschlichen Handlungen heraus zu interpretieren. Das betrifft zum einen die Forschungshandlungen der beteiligten Wissenschafter*innen – dies wurde in einem eigenen propädeutischen teaminternen Workshop zur Technikethik als transdisziplinäre Integration eingelöst. Zum anderen geht es um die empirischen Werturteile der jeweiligen Stakeholder zur Analyse der Situation in Österreich. Ein darüber zu betrachtender und zumindest hinsichtlich der Geltung als universell anzustrebender verbindlicher technikethischer Wertekanon baut auf einem pragmatischen Begriff der Fairness auf.
Im Gegensatz zum schwerwiegenden und mehrdeutigen Konzept der Gerechtigkeit lässt sich ein entsprechend verstandenes Fairnesskonstrukt als „provisorische Moral“ (Hubig 2007) situationsspezifisch umsetzen. Verbindliche Grundlagen werden dabei durch universelle Menschenrechte und Menschenwürde, Gleichheit vor der Ethik und dem Gesetz, Langzeitverantwortung in systemtechnischen Handlungskulturen sowie diverse Prinzipien mittlerer Reichweite gegeben (Respekt humaner Autonomie, Schadensvermeidung, Heilen und Helfen, Transparenz, Inklusion etc.) (HLEG 2019; Beauchamp und Childress 2001; Funk, Frauenberger und Reichl 2020). In diesem Sinne erscheint Fairness, so der Befund, als ein breites wirksames Konzept als Gerechtigkeit, insofern es nicht nur auf abstrakte Werte, sondern Anwendungsrealisierung zielt.
Schließlich erwies sich einmal mehr Technikethik als Verfahrensform, die mehr sein kann als bloße Ethik – wie auch Philosophie einst als Universalwissenschaft (Funk 2021a). Um erstere in methodisch reflektierten Stufen zu erarbeiten, erwies sich der in diesem Abschnitt beschriebene Workshop als ausgesprochen zielführend. Weiterhin diente er dazu, sprach- und vernunftkritische Voraussetzungen gemeinsam zu erarbeiten und zu üben, die in nachfolgenden Gesprächen mit externen Stakeholdern und insbesondere in dem als nächstes beschriebenen Expert*innen-Workshop gewinnbringend eingesetzt werden konnten.

4. Ergebnisse des Expert*innen-Workshops „Faire KI“

Am 22.10.2020 fand unter Leitung von Susanne Wolf-Eberl ein Expert*innen-Workshop statt. Angesichts der Pandemie-Situation wurde dieser Workshop online durchgeführt. Die erzielten Ergebnisse und vertretenen Standpunkte wurden danach in ausgewählten Nachgesprächen weiter geklärt und vertieft.

Die Expert*innen wurden unter dem Gesichtspunkt einer möglichst breiten Sichtweise auf faire KI ausgewählt und vertraten mit Soziologie, Rechtswesen, Mobilitätsforschung, Pflege und Wirtschaftsberatung sowie KI-Forschung und User Experience Design sowohl die Nutzersicht einer breiten Palette von möglichen Anwendungsgebieten als auch die Forschungsperspektive. Eingeladen war auch eine Expertin zu partizipativen Methoden auf digitaler Ebene (ÖGUT), die kurzfristig verhindert war.

Aufgrund der besonderen Online-Randbedingungen wurde außerdem darauf geachtet, die Gruppe nicht allzu groß zu machen, um eine interaktive Diskussion zu ermöglichen.

Da der Workshop aufgrund Covid 19 nicht im Präsenzmodus stattfinden konnte, mussten die ursprünglich intendierten Workshop Methoden den neuen Online-Gegebenheiten angepasst werden. Als Plattform wurde daher Zoom gewählt, und als weiteres Unterstützungstool Mural, wobei allerdings nicht alle Teilnehmer*innen sich in der parallelen Nutzung von Zoom und Mural zurechtfinden und den Schritten - im damals noch ungewohnten Unterstützungstool – folgen konnten. Der Workshop wurde aufgezeichnet und inhaltlich analysiert.

Der Workshop war in drei Hauptteile wie folgt gegliedert:

- Diskussion und gemeinsame Reflexion der TeilnehmerInnen zu den Fragenblöcken, gestützt durch Text- und Bildvorgaben in PowerPoint
- Abgewandelte Stationen-Gespräche zur Kommentierung von Impulsen (Eingangsdefinition von KI, etc.)
- Soziometrische Darstellung zur Sichtbarmachung der Eigenpositionierung zwischen Dystopie und Utopie (samt Darlegung der Begründung)

Inhaltlich wurden dabei vor allem folgende Aspekte beleuchtet:

- Persönlicher Zugang zu fairer KI aus der Perspektive der eigenen Expertise
- Positive und negative Berührungspunkte mit KI
- Stimmungsbarometer zu KI zwischen Dystopie und Utopie
- Stellenwert von partizipativen Möglichkeiten zur Mitgestaltung von KI basierten Systemen

Die Auswertung der Ergebnisse erfolgte in Form einer thematischen Analyse in folgenden Schritten:

- Dokumentation der im WS dargelegten Überlegungen
- Auswertung der dargelegten Meinungen
- Gruppierung der Argumente und Meinungen und Darlegung des Meinungsspektrums

- Analyse und Ableitung von Hauptaspekten zu fairer KI

4.1. Validierung der Eingangsdefinition

Im Rahmen der Workshop-Vorbereitung wurde vom internen Projektteam die in Kapitel 2 beschriebene – bewusst kontrovers formulierte – Definition von KI erarbeitet, um daraus im Rahmen des Workshops Motive für einen kritischen Dialog zu gewinnen. Die Reaktionen der Teilnehmer*innen auf diese Definition waren sehr zurückhaltend. Hinterfragt wurde beispielsweise, was genau die Bedeutung von „bildet Aspekte (menschlicher Intelligenz mit Computersystemen) nach“ sein soll. Welche Aspekte bildet KI nach? Sind es Aspekte menschlicher Intelligenz, die nachgebildet werden oder sind es eher „Schnittstellen und Interfaces zwischen Mensch- und Computersystem, die es zu überbrücken gilt“, wurde hinterfragt. Als Anregung wurden Schnittstellen oder auch die Brückenfunktion zwischen menschlicher Intelligenz und Computersystemen als fehlend konstatiert.

Ein weiterer Schwerpunkt der Diskussion betraf die Zielsetzungskompetenz der KI, der gegenüber eine gewisse Skepsis der Teilnehmer*innen sichtbar wurde, wie sie sich auch im dritten Unterpunkt der Eingangsdefinition widerspiegelt. Echte Autonomie würde demgegenüber ja zum Beispiel auch die selbstständige Setzung und Bearbeitung von Zwischenzielen auf dem Weg zur Erfüllung des eigentlichen Zwecks voraussetzen.

Zudem stellte sich die Frage, inwieweit ein vom Menschen vorgegebenes Ziel für eine KI auch eindeutig funktional übersetzbar ist, also an der Schwelle von menschlicher Alltags- und Fachsprache zur Maschinensprache keine Bedeutungsverluste erleidet. Genau in diesem Bedeutungsverlust liegt ja sehr oft die Quelle der reduktiven Qualität von KI, da eine Übersetzung qualitativer lebensweltlich relevanter Ziele in Maschinensprache zwangsläufig die Notwendigkeit entsprechender Quantisierung nach sich zieht. Allein das ist schon ein sehr schwieriger Punkt, der eine kritische Betrachtung der Chancen und Risiken des zielorientierten Informationsaustauschs zwischen Menschen und KI auf den Plan ruft. Zudem ist zu hinterfragen, ob KI von vornherein auf menschliche Intelligenz zugespitzt werden soll. Exemplarisch hierfür sei die kritische Ergänzung einer Teilnehmerin genannt: Die Formulierung *„KI bildet Aspekte menschlicher Intelligenz nach“* vermisse demnach die Schnittstellenbetrachtung und den Brückenschlag zwischen menschlicher Intelligenz und Computersystemen.

Hinzu tritt die normative Ebene, die sich auch durch unsere Sprache zieht. Allzu häufig verhandeln wir eben nicht in funktional eindeutigen Termini über den ethischen Werterahmen unserer Gesellschaft, sodass einige Fragen thematisiert wurden: Wie kann in der jeweiligen Zielsetzung bereits Fairness als Anforderung mit einfließen? Welche Rolle kann Fairness für die Schaffung von Vertrauen in KI spielen? Kann es Fairness ohne Transparenz geben? Worin unterscheiden sich Fairness und Gerechtigkeit? Sind Misstrauen und Angst vor KI-basierten Systemen ein Ausdruck der Ohnmacht des einzelnen? Ist es sinnvoll, mehr Raum für Verständnis, Mitsprache und Mitgestaltung zu schaffen?

In einem Nachgespräch wurde weiterhin die Operationalisierung von sozialen Konstrukten als „heikler Interpretationsschritt“ angesprochen. Dies gilt insbesondere auch für einen mehrdeutigen Begriff wie Fairness und den Versuch, dies in einen funktional eindeutigen Begriff überzuführen. Besonders angezweifelt wurde, ob soziale Intelligenz, die

stark kulturabhängig ist, nachzubilden ist, beziehungsweise wer hier die moralische und ethische Instanz ist, die über die Güte der Nachbildung und Fairness entscheidet.

4.2. Zugänge zu Fairer KI

Nachstehend sind die Erkenntnisse aus der Perspektive der Teilnehmer*innen dieses Expert*innen-Workshops (mit sehr unterschiedlichen Anknüpfungspunkten an KI) dargestellt. Die unterschiedlichen Zugänge **zu fairer KI** spiegeln naturgemäß auch den jeweils spezifischen beruflichen Hintergrund wider.

1.1.5 Faire KI als nachhaltige und inklusive KI

- *„Für mich ist faire KI eine KI, die alle sozialen Gruppen miteinbezieht, also den Zugang ermöglicht und Diskriminierung vermeidet, aber gleichzeitig auch fair zur Umwelt, zur Natur und zu künftigen Generationen ist. Also für mich ist eine faire KI eine nachhaltige KI.“ (Zitat Workshop-Teilnehmer*in)*[7]

Als „fair zur Umwelt“ wurde in einem Nachgespräch die Priorisierung von ökologischen vor ausschließlich wirtschaftlich-effizienten Algorithmen konkretisiert. Eine mögliche Abwägung zwischen ökonomischen und ökologischen Zielsetzungen sollte nicht unfair - also zu Lasten der Umwelt und des Klimas - ausgehen, da der Klimawandel und falsche Entscheidungen die Lebensgrundlage für zukünftige Generationen beeinflussen und auch Migrationsströme auslösen können. Die KI soll genutzt werden, um eine evidenzbasierte, transparente Datenbasis für Entscheidungssysteme zu erstellen. Fairness ist hier im Sinne einer gleichzeitigen Verfolgung von sozialen, ökologischen und ökonomischen Zielen gemeint. Die Unterstützungsleistung der KI bezieht sich dabei einerseits auf die automatisierte Ergänzung fehlender Daten und andererseits auf der Erstellung eines KI-basierten, fairen „Ausgleichsmodells“. Die SDGs sind wesentlicher Bestandteil dieser globalen und holistischen Betrachtung.

1.1.6 Faire KI als Teil eines Diskurs- und Aushandlungsprozesses der Gesellschaft

Für einen erfolgreichen gesellschaftlichen Diskurs schien den Expert*innen wichtig, einen Schwerpunkt auf erklärbare KI („explainable AI (artificial intelligence)“) sowie Transparenz zu legen. Dies spiegelt sich in mehreren Äußerungen wider:

- *„Eine faire KI ist für mich eine Explainable KI, für mehr Verständnis der Wirkungsweise von KI und zur Nachvollziehbarkeit.“*
- *„Ich stelle die Gegenfrage: Was wäre faire Intelligenz? Die gibt es wahrscheinlich nicht, das ist ein Diskurs, den die Gesellschaft einfach ausverhandeln muss. Je nachdem worauf sich die Gesellschaft einigt, was fair ist, das wäre dann eine faire KI. Explainable AI ist sicherlich ein wichtiger Aspekt, ebenso wie Transparenz. Um festzustellen, was Fairness ist, müssen die Leute auch verstehen, um was es geht und wie der Algorithmus entscheidet.“*

[7] Wie hier werden, sofern nicht anders markiert, im gesamten Rest von Kapitel 4 transkribierte Originalzitate der Workshop-Teilnehmer*innen kursiv und eingerückt gekennzeichnet.

1.1.7 Faire KI als rechtliche Möglichkeit

In einem gewissen Sinne könnte das Konzept einer fairen KI auch die Einbettung in einen entsprechenden Rechtsrahmen beinhalten, der es ermöglicht, Diskriminierung zu vermeiden oder gegen fehlerhafte KI-basierte Entscheidungen vorzugehen.

- *„Die Möglichkeit zu seinem Recht zu kommen, die Vorsehung eines Rechtswegs und einer Schadenersatzmöglichkeit."*

Hierbei ist zu bedenken, dass die Frage nach möglichem Schadensersatz zunächst eine entsprechende gesetzliche Regelung des Schuldprinzips voraussetzt, die ja heute noch weitgehend ungeklärt ist.

1.1.8 Faire KI als Transparenz von und Verantwortlichkeit für Algorithmen

Datengüte benötigt entsprechende Überprüfungsmechanismen, besonders in mehrstufigen Datenverarbeitungsprozessen. Die Vermeidung von *Biases* und Fehlern gründet sich bei fairer KI zum einen auf qualitativ hochwertigen Trainings- und Inputdaten und wird zum anderen unterstützt durch Transparenz und Rechenschaftspflicht bzw. Verantwortung.

- *„Also wenn man sich den ganz bekannten AMS Algorithmus ansieht, der nicht ganz zum Einsatz gekommen ist. Da hat man – rein durch die Klassifizierungsmerkmale - schon gemerkt, dass da Klassifizierungen drin sind, die nicht so fair ablaufen. Ja, aber das bildet ja nur unsere Gesellschaft ab. Ich glaube schon, dass ganz viele Sachen durch zivilgesellschaftliche Organisationen/Vereine/Initiativen aufgezeigt werden."*
- *„Die Transparenz und die Nachvollziehbarkeit sind ganz große Probleme. Dass da so eine Blackbox ist und das dann aber als Entscheidung oft nicht hinterfragt wird."*
- *„Entscheidungstreffende Algorithmen können das eigene Leben schon auf negative Art und Weise beeinflussen, wo man nicht wirklich nachvollziehen kann, was das System eigentlich macht und es hat - gerade in sensiblen Bereichen, wo es um Jobsuche geht, wo es um Menschen vor Gericht oder in Gefängnissen geht- brutale Auswirkungen. Das wird dann selten hinterfragt."*
- *„Ein positiver Mehrwert von KI wäre, wenn man sich sozusagen auch nutzerseitig darauf verlassen kann, dass das ein korrekter, aktueller Datensatz ist, der da dahintersteckt."*

1.1.9 Faire KI als Beitrag zu besseren Lebensumständen

Schließlich wurde von den Expert*innen betont, dass fairer KI jedenfalls auch großes Potential zugesprochen wird, das alltägliche Leben in zentralen Bereichen signifikant zu verbessern:

- *„AI kann in den meisten Bereichen etwas Positives leisten, also ob das jetzt die logistische Versorgung mit Lebensmitteln, Wirtschaftswachstum durch Effizienzsteigerung oder medizinische Versorgung ist".*

Zusammenfassend bleibt festzuhalten, dass eine der Ausgangsannahmen des Projekts darin bestand, dass die fehlende Einbeziehung von Personen bei der

Technologieentwicklung negative Folgen für die betroffenen Individuen nach sich ziehen kann. Um zum Vertrauen in technologische Systeme beizutragen, erscheint es wesentlich, positive wie negative Berührungspunkte der Expert*innen mit KI-basierten Technologien zu verstehen. Diese spiegeln die vielfältigen Erwartungen, Hoffnungen und Mythen wieder und tragen dazu bei, den Fairnessanspruch an KI zu konkretisieren und die relevanten Hürden zu erkennen. Positive Berührungspunkte und eine positive Stimmungslage gegenüber KI-basierten Entwicklungen können unternehmensseitig genutzt werden, um Barrieren und Vorurteilen entgegenzuwirken. Negative Berührungspunkte zeigen die Sensibilisierung auf mögliche Problembereiche und Hürden auf. Gemeinsam mit der eigenen Verortung zwischen Dystopie und Utopie wird die Stimmungslage ausgelotet.

4.3. Positive Erwartungen an eine Faire KI

Als positive Erwartungen an eine faire KI lassen sich aus Sicht der Expert*innen drei Hauptaspekte zusammenfassen:

- Stärkung der Wirtschaftskraft
- Veränderung in Richtung einer holistischen Betrachtungsweise
- Neue Nutzungs- und Leistungsversprechen durch KI-basierte Lösungen

1.1.10 Stärkung der Wirtschaftskraft

- Die Potenziale der KI liegen in gesteigerter Produktivität und Effizienz. Insbesondere der effizientere Energieeinsatz aufgrund KI-basierter Systeme kann zu einer Entlastung der CO_2-Belastung führen. Die KI unterstützt die Steigerung der Wirtschaftsleistung und eröffnet neue Möglichkeiten für verschiedenste Wirtschaftszweige.
- Big Data und Bewegungsdaten von Smartphones sind die Basis für viele KI-basierte Anwendungen. Hierdurch besteht eine große Aufbruchstimmung bei Unternehmen etwa im Mobilitäts- und Energiesektor. Selbstlernende Systeme und Methoden der KI sind wesentliche Antriebsfaktoren der digitalen Transformation, wobei kollaborative Zusammenschlüsse von Unternehmen die Basis für neue KI-basierte Geschäftsmodelle mit Hilfe von geteilten Daten und neuen KI-Technologien bilden. Dabei stehen Echtzeitdaten, Bewegungsdaten (Mobilität) und Smart-Meter-Daten (Energie), die für KI-basierte Services genutzt werden, im Zentrum.
- Eine starke Hoffnung liegt auf den selbstregulatorischen Kräften der Marktwirtschaft. Wenn der Markt die Nutzungsbedingungen im Web regelt, bedürfte es demzufolge idealiter keiner regulatorischen Eingriffe, um die Rechte der Nutzer*innen zu schützen.
  - ➢ *„Im privatrechtlichen Bereich wird schon noch ein bisschen die Autonomie des Marktes gewahrt werden müssen, denn sonst wird es kaum zu neuen Entwicklungen kommen.“*

Allerdings gab es hierzu auch konträre Meinungen, die die nicht funktionierenden Selbstregulierungssysteme der großen Tech Giganten und Fake News angesprochen haben und nach einer staatlichen Regulierung verlangt haben. Weiterhin wurde angemerkt,

dass Ausnahmezustände wie die weltweite Corona Pandemie Zulassungen von bildbasierten Diagnosesystemen im Web beschleunigen, die die Basis für KI-basierter Telemedizin und Forschung als neues Geschäftsfeld sind.

Insgesamt gesehen waren die Diskutant*innen mehrheitlich der Meinung, dass faire KI im Interesse aller Unternehmen liegt. KI eröffnet Marktchancen in ungeahntem Ausmaß, mit neuen Geschäftsmodellen auf Basis von Daten und ständig (weiter)entwickelten Methoden der Künstlichen Intelligenz. Dabei wird von starken Selbstregulationsmechanismen des Markts ausgegangen, der schnell reagiert. Insofern wird KI als eine Entwicklung wie viele andere auch gesehen, wobei Fairness-Aspekte wesentlich in der Verantwortung der Unternehmen liegen. Allerdings erfordern KI-basierte Lösungen unterschiedliche Kernkompetenzen, die ein Unternehmen (und hier vor allem KMUs) kaum allein abdecken kann. Daher scheinen Fairnessaspekte im Hinblick auf kleinteilige Verantwortlichkeiten und komplexe Datenstrukturen oftmals schwierig zu beurteilen.

1.1.11 Veränderung in Richtung einer holistischen Betrachtungsweise

- Einige Diskutant*innen waren der Meinung, dass KI-basierte Systeme die globalen Problembereiche und Herausforderungen der Menschheit lösen werden. Demzufolge werde KI die Welt verbessern, indem einerseits innovative, technologische Optimierungsansätze und selbstlernende Systeme zum Tragen kommen, die die Knappheit von globalen Ressourcen sowie unternehmerische Risiken bewerten und einkalkulieren, und andererseits ökonomische, soziale und ökologische Aspekte als gleichwertige Zielgrößen gesehen werden.
- Grundsätzlich wurde von den Expert*innen anerkannt, dass eine Faire KI einen positiven Beitrag zu den *Sustainable Development Goals* (SDGs)[8] liefert.
- Gewünscht sind aus Expert*innen Sicht:
  - *„vermehrt interdisziplinäre Teams (anstatt von ‚techniklastigen Teams'), um mit Lösungsansätzen besser den Bedürfnissen verschiedener Personengruppen zu entsprechen".*

Auch wenn interdisziplinäre Teams vielfach gewünscht sind, sieht die Praxis noch sehr IT-lastig und technikdominiert aus. Die Einbeziehung insbesondere der Ethik und der Diskurs mit Betroffenen und Vermittler*innen ist ein erster Schritt zu einer gesamtheitlichen Betrachtung.

Insgesamt betrachtet gab es in Bezug auf das Potential von KI eine sehr optimistische Stimmungslage unter den Expert*innen. Das Vertrauen, dass sich KI-basierte Lösungsansätze auch ausgleichend auf bestehende Ungerechtigkeiten und Ungleichheiten auswirken, war größtenteils vorhanden. Mehrheitlich einig war man sich, dass eine starke politische Kraft regulierend eingreifen sollte, wenn es zu offensichtlichen Benachteiligungen und Diskriminierungen kommt (allerdings wurde diese Fragestellung nicht vertieft).

[8] https://sdgs.un.org/goals

Daher scheinen Methoden der künstlichen Intelligenz, die auf globalen Daten basieren und damit auch globale Optimierungsaufgaben lösen könnten, durchaus wünschenswert. Wenn inter- und transdisziplinäre Teams nutzer*innenorientiert arbeiten und unterschiedliche Fairnessaspekte einbringen, dann wäre ein Expert*innensystem, das Entscheidungsregeln klar strukturiert und als Entscheidungsalgorithmus transparent und nachvollziehbar darstellt wünschenswert.

1.1.12 Neue Nutzungs- und Leistungsversprechen durch KI-basierte Lösungen

- Aus Sicht der Diskutant*innen bestehen positive Entwicklungsmöglichkeiten und Einsatzmöglichkeiten von KI im Mobilitätssektor beispielsweise bei Fahrassistenzsysteme, die aus über Sensorik generierten Daten neue Erkenntnisse in Echtzeit und Infos liefern und damit zunehmend den Weg zum autonomen Fahren ebnen. Weitere wichtige KI-basierte Nutzungsversprechen liegen im Sicherheitsgewinn (beispielsweise im Bereich *Ambient Assisted Living* bei bewegungsdatenbasierter Sturzprävention). Ferner gelten neue KI-basierte Unterstützungstechnologien wie der Pflegeroboter als Hoffnungsträger in der Altenversorgung.
- Die höhere Verlässlichkeit von bspw. Echtzeitdaten wird als neues Leistungsversprechen gesehen. Prognosedatenmodelle auf Basis von Echtzeitdaten ermöglichen beispielsweise genaue Stauvorhersagen und Zeitprognosen; auch der öffentliche Personenverkehr sowie die Logistikbranche nutzen Echtzeitdaten für neue Leistungsversprechen (Ankunftsvorhersagen).
- Ein anderes Leistungsversprechen betrifft die gleichmäßige Arbeitsleistung der Maschine im Vergleich zu menschlicher Arbeitskraft. Ein Pflegeroboter etwa ermüdet nicht, vergisst keine Arbeitsschritte und gilt damit als verlässlicher. Als weiteres Beispiel genannt wurde ein sensor-basiertes Kontrollsystem, das reagiert und Alarm schlägt, wenn Demenzkranke einen definierten Radius verlassen. Erwünscht ist also eine*r Teilnehmer*in zufolge eine

  ➢ *„KI, die auf mich aufpasst, wenn ich schlafe – der elektronische Freund und Aufpasser.“*

Die virtuelle Krankheitsdiagnostik auf Basis von Mustererkennung und Bilderkennung wird als zukunftsweisende Entwicklung für Regionen mit geringer Arztdichte gesehen. Es werden Bilderdatenbanken zu Krankheiten und Verläufen aufgebaut werden, die gemeinschaftlich genutzt und ausgewertet werden können. Hier entstehen neue KI-basierte Leistungsangebote.
Ein anderes Leistungsversprechen zielt auf Nischenprodukte ab, die hohen Nutzen für Betroffene stiften:

➢ *„Vergessene Personengruppen in einer Gesellschaft, wie Kleinstgruppen, die scheinbar nicht profitabel genug waren, um ihre Probleme und Krankheiten zu lösen haben nun Dank KI-basierten Methoden und Verfahren eine neue Chance. Z.B. Querschnittsgelähmte, die sich wieder bewegen können oder Personen, die sich mittels Augensteuerung verständigen können“.*

Den Expert*innen zufolge gibt es also eine Fülle konkreter Nutzungs- und Leistungsversprechen, die als KI- basierten Dienstleistungsangebote oder Produkte auf den Markt

kommen, um die Lebens -und Arbeitsqualität verbessern. Diese sichtbaren Innovationen sind es, die das Verständnis und Vertrauen in die KI heben.

Insbesondere besteht beispielsweise die Hoffnung und Erwartung, dass Bilderkennungssysteme und die zugrundeliegenden Bilddatenbanken immer vollständiger und besser werden. Bilderkennung ist dabei in vielen Fällen die Grundlage für autonomes Verhaltens von Maschinen und von Videoüberwachung. Auch in der Medizin (bspw. Hautkrebs/Tumorerkennung) wird Bilderkennungssoftware eingesetzt. Die Technologie in der Bilderkennung ist noch nicht ansatzweise perfekt, zum Beispiel arbeitet gängige Gesichtserkennungssoftware weder fehler- noch diskriminierungsfrei. Andererseits ist der Nutzen klar erkennbar, und weitere Investitionen in Pilotprojekte und dem Aufspüren von Fehlern wären demzufolge sinnvoll.

4.4. Negative Berührungspunkte in Bezug auf faire KI

Negative Berührungspunkte der Expert*innen in Bezug auf faire KI zeigen die unterschiedlichen Facetten des Problembewusstseins auf. Hierbei wurden folgende vier Aspekte identifiziert:

- Mangelndes Vertrauen und Ohnmacht
- Digitale Spaltung der Gesellschaft
- Mangelnde Datenqualität
- Unklare Regulierung

1.1.13 Mangelndes Vertrauen und Ohnmacht im Fall von Fehlentscheidungen und Diskriminierung

Nach Ansicht der Expert*innen ist ein Verlust der Selbstbestimmtheit durch ständige Kontrollmechanismen und damit Angst vor Überwachung und falschen Schlussfolgerungen denkbar:

> ➢ *„Bei Connect Care ging es um Aktivität und Inaktivität im Raum, mit Meldung an die Familie mit Infos wie: ist schon aufgestanden, oder: der Herd ist an, wo man aus der Ferne auch den Herd abschalten konnte. Da war dann die Angst vor Überwachung stark, und wer weiß wer das alles sieht und Schlüsse daraus zieht.“*
>
> ➢ *„Ein Beispiel, das mich immer so ein bisschen beschäftigt, sind Sturzsensoren in Privathaushalten von Senior*innen. Da gab es vor ein paar Jahren so ein Paper, das gezeigt hat, dass hochaltrige Menschen ein romantisches Tête-à-Tête hatten, und der Sturzsensor hat das dann als außergewöhnlich, quasi als Notfall, erkannt. Deswegen beschäftigen mich diese fehlerhaften Kategorisierungen.“*

- Ein besonderes Problem wird im Fehlen menschlicher Ansprechpartner*innen bei automatisierten Entscheidungen gesehen. Gerade bei KI-basierten Entscheidungen, die diskriminierend, ungerechtfertigt und fehlerhaft sind, ist es besonders einschneidend, wenn sich niemand zuständig fühlt und es keine vorgesehene Beschwerdestelle bzw. einen Instanzenweg gibt, denn gerade benachteiligte Gruppen scheuen kostenpflichtige Einspruchsmöglichkeiten und langwierige Prozesse.

- *"... ein Kampf gegen Windmühlen..."*

- Wenn einer KI Entscheidungen mit weitreichenden Konsequenzen übertragen werden, dann kann das aufgrund des potentiellen Korrekturaufwands zu deren praktischer Irreversibilität führen:
  - *„Hoher Aufwand, um Fehlentscheidungen zu berichtigen und eigentlich keine Möglichkeit falsche Schlüsse zu unterbinden."*
  - *„Bei Einspruch stößt man gegen eine Mauer. Also man hat ihr aufgrund ihres Widerspruchs eine menschliche Zwischenebene eingerichtet, mit der sie sprechen konnte, die sich aber nur auf dieses (fehlerhafte) System berufen hat und das System auch nicht in Frage gestellt hat."*

- Vor allem auf Verbraucherseite finden sich Ängste, in sensiblen Bereichen mehr oder weniger ausgeliefert zu sein. Zum Beispiel ist der Online-Handel und die Nutzung von digitalen Angeboten gelegentlich mit dem Unbehagen verbunden, dass die Privatsphäre nicht ausreichend geschützt ist und zu viele unzulässige personenbezogene Daten gespeichert und für KI-basierte Optimierungsentscheidungen verwendet werden, was bis hin zu manipulativem Verhalten reichen kann. Es ist unklar, aus welchem User-Verhalten (Spuren im Netz) welche Schlussfolgerungen gezogen und missbräuchlich verwendet werden können. Zudem fühlen sich Internetuser*innen vielfach beobachtet und nicht in der Lage, sich dagegen zu wehren. Es ist unklar und undurchsichtig worauf Suchprofile zugreifen und welche Informationen gespeichert und verknüpft werden.
  - *„Ein System wie Google, das einen relativ stark einschränkt und immer wieder dieselben Sachen anzeigt, das einen trackt, auch wenn man die Google Suchmaschine nicht mehr verwendet."*

1.1.14 Digitale Spaltung der Gesellschaft: Mangelndes Verständnis, mangelnde Einbeziehung von Nutzer*innengruppen, mangelnde Berücksichtigung des Faktors Mensch

Unter den Expert*innen wurde diskutiert, ob und wie mit der raschen Digitalisierung Personen ohne Internetzugang und Smartphone den Anschluss an die Gesellschaft verlieren können. Diese Problematik geht oft einerseits mit sozialer Differenzierung, andererseits mit dem Alter einher. Das mangelnde Verständnis dafür, wie KI-basierte Angebote und Lösungen funktionieren, löst bei Personen mit niedrigem Bildungsgrad oftmals Misstrauen, Unmut und Verärgerung aus. Nutzenstiftende KI-basierte Lösungen sind aber ohne entsprechende Akzeptanz und einem klaren Verständnis der Vorteile durch eine Zielgruppe zum Scheitern verurteilt. Die Ablehnung KI-basierter Lösungen – auch wenn sie die eigene Unabhängigkeit unterstützen würden – resultiert oft aus dem Unverständnis für deren technologische Wirkungsweise.

Als problematisch erweist sich aus der Perspektive der Expert*innen insbesondere die Ablehnung KI-basierter Hilfsangebote in der Pflege durch teilweise wenig technikaffines Pflegepersonal, das einerseits in den KI-basierten Lösungsansätzen die Gefahr für den eigenen Arbeitsplatz sehen und andererseits eine potentielle Überwachung ihrer Tätigkeit und Einschränkung ihres Ermessensspielraums orten. So löst in diesem Fall die mangelnde Einbindung von Betroffenen und Betreuer*innen in die technologische Entwicklung von Assistenzlösungen Abwehrreaktionen und Verunsicherung aus, zudem

führt mangelndes Verständnis für die permanente Datenerfassung und die Wirkungsweise zu Irritationen.

- *„Vermittler*innen / Anwender*innen von KI brauchen eine höhere Wertigkeit."*
- *„KI soll den Menschen nicht ersetzen - das Personal braucht die Sicherheit, nicht den Job zu verlieren."*
- *„Es gibt keine Nutzer*innen-Zentriertheit."*

1.1.15 Mangelhafte Datenqualität und fehlende Methoden

Ein für die Expert*innen zentraler Aspekt betrifft die Qualität der Daten, auf denen KI-Anwendungen aufbauen, insbesondere im Hinblick auf fehlerhafte Kategorisierungen, Fehlerquellen und möglicher Bias. Bemängelt wird u.a. die oft fehlende Transparenz der Datenquellen, welche sich oft als Black Box darstellen, und eine daraus resultierende mangelnde Nachvollziehbarkeit. Hier fehlen nutzer*innen-zentrierte Designmethoden und *Explainable AI*.

- *„Die KI wird vom Menschen entwickelt und kann dadurch niemals fehlerfrei sein."*
- *"In Bezug auf faire KI würde ich einmal da ansetzen, wo wir noch gar nicht beim Entwickeln sind, also bei den Anforderungen an Daten, beim Datenfinden. Bevor überhaupt faire KI entsteht, würde ich bei einer fairen KI auch nie den Begriff „dark patterns" vergessen, weil es da auch in ganz vielen Entwicklungen um designte Muster geht, die von vornherein unfair sind, wenn man so will. Dark Pattern beschäftigt eigentlich mittlerweile auch das Rechtssystem, da sollten wir auch ansetzen, wenn es um Unfairness geht."*

1.1.16 Unklare Regulierung

Auf dem Gebiet der Regulierung stellen sich den Expert*innen viele offene Fragen, welche von potentiellen Rechtsbrüchen durch Verletzung der Grundrechte (z.B. bei KI-basierten Gesichtserkennungssystemen) bis hin zu ungeklärten Aspekten der Produkthaftung reichen. Die Hoffnung auf eine hinreichende Selbstregulierung innerhalb des Marktes stößt zumindest im Hinblick auf die großen Konzerne (Facebook, Google, etc.) auf Skepsis. Schließlich ist aufgrund der derzeitigen Rechtslage auch eine mangelnde Nutzer*innenzentriertheit und Barrierefreiheit zu konstatieren.

Insgesamt sind sich die Expert*innen vieler Vorbehalte, Risiken und Schwachstellen bewusst, die sich negativ auf die Vertrauenswürdigkeit von KI-basierten Systemen und Angeboten auswirken können. Es ist für sie weitgehend klar, dass eingeforderte Fairnessmaßnahmen (wie bspw. mehr Transparenz und Nachvollziehbarkeit) unerlässlich, aber gerade bei KI-basierten Systemen oftmals schwierig realisierbar sind. Daraus lässt sich ableiten, dass eine umfassende Verpflichtung auf Grundsätze der Explainable KI das Vertrauen der Bürger*innen entscheidend stärken würde.

- *„Unter fairer KI kann im besten Falle vieles passieren, dass bisher in der HCI Entwicklung gefehlt hat bzw. einfach nicht gemacht wurde: Design Research, Scenario based Design, Evidence based Design, Nutzerzentriertheit, Accessibility, Diversität, Stakeholder Vielfalt, Nachhaltigkeit etc."*

Gewünscht wurde von einigen Expert*innen ein starker Staat, der die Regulative vorgibt, da man insbesondere im Fall der KI

- *„die staatliche Verantwortung nicht der Zivilgesellschaft überlassen kann.“*

4.5. Zusammenfassung und Schlussfolgerungen

Insgesamt hat der Expert*innenworkshop seine Ziele vollumfänglich erreicht. Die Diskussion der bewusst kontrovers formulierten Definition von KI förderte wichtige Aspekte zutage, wie etwa Fragen der Zielsetzungskompetenz von KI, der Möglichkeit einer funktionalen Übersetzung von Zielen wie auch der Operationalisierung zugehöriger sozialer Konstrukte im Allgemeinen.
Bei der Erörterung möglicher Zugänge zu fairer KI stellte sich heraus, dass zunächst einmal Nachhaltigkeit und Inklusion eine wichtige Rolle spielen und Fairness im Sinne der gleichzeitigen Verfolgung von sozialen, ökologischen und ökonomischen Zielen verstanden wurde. Sodann wurde von einigen Expert*innen die Notwendigkeit eines gesellschaftlichen Aushandlungsprozesses betont, um das Ziel einer „Explainable KI“ und damit Transparenz zu erreichen. In diesem Zusammenhang wurde dann auch die Frage nach Verantwortung für mögliche Diskriminierungen und *Biases* thematisiert und von einigen Expert*innen die Schaffung eines entsprechenden rechtlichen Rahmens gefordert. Einer so verstandenen und auf öffentlichem Diskurs basierenden fairen KI wurde schließlich das Potential attestiert, zu einer Verbesserung genereller Lebensumstände signifikant beizutragen.
Im weiteren Verlauf des Workshops wurden sodann positive Erwartungen der Expert*innen an eine faire KI entsprechenden negativen Berührungspunkten damit gegenübergestellt. KI im Allgemeinen und eine faire KI im Besonderen eröffnet demnach große wirtschaftliche Möglichkeiten auf vielen Gebieten, wobei insbesondere Aspekte einer nachhaltigen (wie etwa bezüglich der CO2-Problematik) und holistischen Betrachtung erwähnt wurden. Unterstrichen wurde auch die Hoffnung auf selbstregulatorische Kräfte innerhalb der Marktwirtschaft, die das Streben nach Effizienz- und Produktivitätsfortschritten in die richtigen und vor allem wünschenswerten Bahnen lenkt, um so die Leistungsversprechen, die mit KI verbunden werden, auf gesellschaftlich akzeptable Weise zu realisieren.
Demgegenüber zeigte sich aber auch von einigen Expert*innen eine gewisse Skepsis, wenn es um die Frage der Vermeidung von Diskriminierung geht, ja man fürchtet sogar eine gewisse Ohnmacht im Falle von Fehlentscheidungen, die schwer revidierbar sind. Auch mangelnde Datenqualität stellt sich als mögliche Quelle von Ängsten heraus, sodass sich letztlich die Möglichkeit einer gesellschaftlichen Spaltung nicht ausschließen lässt. Daher bedarf es nach Meinung mancher Expert*innen trotz allem einer starken Regulierung, was natürlich in einem gewissen Widerspruch zur Hoffnung auf Selbstregulation des Marktes steht.
Damit also faire KI die damit verbundenen Erwartungen an wirtschaftlichen Erfolg wie auch die Verbesserung von Umwelt und Lebenswelt einlösen kann, stellt sich das Streben nach Erklärbarkeit (*explainability*) und die Notwendigkeit eines darauf aufbauenden umfassenden gesellschaftlichen Diskurses als zentral heraus. Dies lässt sich einerseits durch die massive Verstärkung eines geeigneten Wissenstransfers einerseits von Entwicklern und Anbietern von KI-Technologie zu den Nutzern, und andererseits von Personen mit viel Erfahrung in der Arbeitsweise von KI-basierten Systemen zu Personen mit

wenig Kenntnissen und vielen Ängsten vor KI erreichen, um auf diese Weise das grundlegende Verständnisse für KI-basierte Systeme zu steigern und entsprechende Ängste nach Möglichkeit zu relativieren. Andererseits stellen partizipative Austauschformate zusammen mit geeigneten Methoden zur Erklärung, Visualisierung, Entmystifizierung, Klarstellung, aber auch In-Fragestellung von KI und KI-basierten Technologien einen vielversprechenden Ansatz dar, um die hierfür notwendigen Diskursräume zu schaffen. Methodisch bieten sich hierfür partizipative Ansätze der Mitgestaltung wie auch neue Methoden aus dem Design Research und aus dem nutzer*innen-zentrierten Design wie auch geeignete Peer-to-Peer Formate an, um Räume zur Aushandlung von „Orientierungshilfen“ zur Verfügung zu stellen, in denen beispielsweise vertrauenswürdige KI-Systeme identifiziert und gekennzeichnet werden und sich andererseits Risiken, wie beispielsweise mögliche Manipulationsgefahr und Missbrauchsgefahr, signalisieren lassen. Schließlich ergibt sich aus dem von den Expert*innen ebenfalls geäußerten Bedürfnis nach einem geeigneten Rechtsrahmen die Notwendigkeit zumindest einer öffentlichen Anlaufstelle für Fragen und Beschwerden, Aufdecken von Diskriminierung sowie das Hinterfragen von Handlungsanweisungen.

5. Herausforderungen bei der Umsetzung

5.1. Vorbemerkungen

Um zu möglichen Umsetzungsstrategien zu gelangen, wurde im Einklang mit den Projektzielen insbesondere das Potential von partizipativen Formaten untersucht. Hierzu liefert die österreichische Gesellschaft für Umwelt und Technik (ÖGUT) in ihrem Methodenhandbuch „Bürgerbeteiligung in der Praxis“ (Handler 2018) einen guten Überblick zu Qualitätskriterien, Einsatzmöglichkeiten, Eignung und Gestaltungsmerkmalen von Beteiligungsmöglichkeiten. Besonders gelungen erscheint eine Übersicht von Methoden im Hinblick auf das Potenzial. Hier werden sechs Stufen definiert: vom Informieren, Meinung/Reaktion einholen, Aktivieren und Diskussionen starten, gemeinsam planen und entwickeln zu längerfristiger Zusammenarbeit und der Konflikte Bearbeitung. Die Methoden reichen beispielsweise von Aktivierender Befragung bis zu Online-Dialog, Open Space und Zukunftswerkstatt (vgl. Handler 2018, S.20) und werden in Bezug auf ihre Eignung bezüglich der sechs Stufen bewertet. Dabei galt eine Kombination aus Online- und Offline-Angeboten in Österreich bisher als besonders erfolgsversprechend, da hiermit unterschiedliche Zielgruppen eingebunden werden konnten. Zukünftig sind Online-Beteiligungsangebote vermutlich leichter zu realisieren, da die Akzeptanz von Online-Tools und Methoden durch den Lockdown in bestimmten Zielgruppen gestiegen scheint.
Allerdings stehen hierbei Unternehmen wie Gesellschaft vor signifikanten Herausforderungen, welche innerhalb des Projektteams diskutiert wurden und die in den folgenden Abschnitten beschrieben werden.

5.2. Herausforderungen aus Unternehmensperspektive

Immer mehr Unternehmen erkennen und nutzen das große Potential von Daten für KI-basierte Produkte und Dienstleistungen. Daraus ergeben sich auch neue Prozessabläufe und Anforderungen an Fairness. Im Hinblick auf interne Abläufe scheint der Effizienzgedanke im Vordergrund. KI-basierte Systeme stoßen, wie im Expert*innen Workshop thematisiert, auf Widerstand bei Mitarbeiter*innen, wenn diese ihren Job gefährdet oder ihren Handlungs- und Ermessensspielraum eingeschränkt sehen.

Die Nutzung von KI-basierten Daten, um eigene KI-basierte Produkte oder Dienstleistungen zu erstellen, ermöglicht gerade neue Geschäftsfelder und daraus entstehen neue Geschäftsmodelle. Die Sensibilisierung auf eine faire KI und die Vermeidung von Diskriminierung ist je nach Branche unterschiedlich.

Je nach Erwartungen, welche Aufgabe der KI zukommt, erscheinen unterschiedliche Aspekte relevant:

- Die Unterscheidung zwischen unternehmenseigener (interner) Implementierung von KI-Systemen oder die (externe) Nutzung von „KI-as-a-Service“-Angeboten
- Die Zugänglichkeit von unterschiedlichen Datenquellen
    - Abschätzung und Vergleichbarkeit der Datengüte und Verwendbarkeit für eigene Zwecke
- Die Verantwortungsübernahme im Umgang mit „geteilten“ Daten

  - Transparenz/Überprüfbarkeit der Herkunft und der Qualität von Daten
  - Entwicklung und Einhaltung eines anerkannten Qualitätsstandards
  - Kontinuierliche transparente Aktualisierung der Datenbasis (Einpflegen von Änderungen, Glättungen, Berichtigungen, Ergänzungen wie digitale Zwillinge etc.)
  - Kennzeichnung von Data-Cleaning Prozessen und Datenselektionskriterien
  - Verantwortungsfestlegung in nachgelagerten Prozessen
  - Datenhoheit
- Die Einhaltung von Verordnungen und Empfehlungen im Hinblick auf vertrauenswürdige KI
  - Schutz der Privatsphäre und Grundrechte
  - Neue Bestimmungen und Empfehlungen (digitale Barrierefreiheit etc.)
- Neue Partnerschaften und Kooperationen
- Neue Geschäftsmodelle
- Diskriminierungsüberprüfung bei KI-basierten Dienstleistungen wie Empfehlungssystemen, Verbraucherscorings, etc.
  - Orientierungshilfen wie Stereotypen und Stigmatisierung entgegnet werden kann
  - Sichtbarmachung von Ungleichheit und Unausgewogenheit

Vertrauensfördernd erscheint somit:

- Die Schaffung eines guten – breit akzeptierten – Qualitätsstandards im Umgang mit Daten
- Die Offenlegung von Qualitätsmechanismen und die Übernahme von Verantwortung im Falle von Problemen
- Ein breiter Diskurs zu möglichen Nachteilen und Benachteiligungen bei Einführung neuer KI-basierter Technologien mit Mitarbeiter*innen, Partner*innen und Nutzer*innen
- Eine Ansprechstelle und ein offenes Ohr für die Klärung von Diskriminierungsverdachtsmomenten

In diesem Zusammenhang sei auf die deutsche Plattform „Lernende Systeme“ hingewiesen, die für einen breiten Dialog und Diskurs wirbt[9]. Diese macht einerseits Unternehmen Mut, ihre Chancen zu erkennen und adressiert, und beleuchtet andererseits bewusst potentielle KI-Mythen, um das Vertrauen in die Möglichkeiten dieser Technologie zu festigen. Eine wesentliche Frage ist es hierbei herauszufinden, wie die KI helfen kann, Ziele zu erreichen. Es wird hier empfohlen, nicht ausschließlich auf interne Prozesse und Effizienzsteigerung zu fokussieren, sondern auch Kundenvorteile (z.B. durch verbesserte Kundenempfehlungssysteme) zu betrachten. Hier ist allerdings kritisch anzumerken,

[9] https://www.plattform-lernende-systeme.de/selbstverstaendnis.html

dass nur ein verantwortungsvoller, transparenter und fairer Kundenempfehlungsprozess langfristig den Nutzen für ein Unternehmen erhöht. Verdachtsmomente in Richtung manipulativer Empfehlungssysteme untergraben das Vertrauen und können nur durch Transparenz und Offenheit entkräftet werden.

5.3. Herausforderungen aus gesellschaftlicher Perspektive

Wenn, wie in dem Expert*innen-Workshop explizit gefordert, ein Aushandlungsprozess der Zivilgesellschaft stattfinden soll, dann entspricht das dem demokratischen Verständnis einer Gesellschaft, welche im Diskurs notwendige Rahmenbedingungen, Regeln und Standards in Bezug auf eine transparente, diskriminierungsfreie KI herstellt. Der Aushandlungsvorgang bewirkt insbesondere eine Stärkung des Vertrauens in das System und erhöht die Bereitschaft, sich an Regeln und Qualitätsstandards zu halten. Dadurch wird faires Handeln als intrinsische Motivation und moralische Verpflichtung (sich an Vereinbarungen des Aushandlungsprozesses zu halten) gesehen.

Partizipative Prozesse involvieren bewusst breite Teile der Bevölkerung und unterschiedliche Stakeholdergruppen. Es sollte jedoch nicht erwartet werden, dass sich Bürger*innen aktiv einbringen, vielmehr müssen Räume und Möglichkeiten geschaffen werden, um Diskurse und Kommunikation stattfinden zu lassen. Hinsichtlich der Teilnahme würde idealerweise ein Random Sampling (Zufallsauswahl) zum Einsatz kommen, dass jeder Person, die gleiche Chance einräumt, Teil des Prozesses zu sein. Ausgewählte Personen sollten die für sie notwendigen Rahmenbedingungen für ihre wichtige Aufgabe erhalten und mit Fakten und Know-How zu KI-basierten Systemen versorgt werden. Somit würden bestehende Ungleichheiten in der Einbeziehung von benachteiligten Gruppen verringert werden, und eine repräsentative Personenstichprobe würde die gesellschaftliche Verhandlung als öffentliche Vertretung wahrnehmen.

In einem weiteren Schritt treten Florian Eyert und Paola Lopez dafür ein, dass „Fairness demokratisch als Aushandlungsprozess über Gerechtigkeit gedacht werden muss und Transparenz als kommunikative Voraussetzung dafür“ (Eyert & Lopez 2021). Dieses Transparenzverständnis beinhalte, dass Orte geschaffen werden, an denen Kommunikation erfolgen kann, was eine öffentliche Aufgabe darstellt. Die beiden weisen darauf hin, dass die Städte Helsinki und Amsterdam öffentlich einsehbare Register mit verwendeten KI-Systemen eingeführt haben, was als erster Schritt zu werten ist. Sie fordern auch eine aktive Institutionalisierung einer Offenheit gegenüber zivilgesellschaftlicher Kritik, die auch darin münden könne, bestehende KI-basierte Systeme abzuschaffen oder gar nicht einzusetzen.

Aus weiteren Literaturrecherchen und bisherigen Erfahrungen mit partizipativen Prozessen erscheint somit als vertrauensfördernd für derartige Prozesse:

- eine klar definierte Zielsetzung
- ein stimmiges Prozessdesign, im Sinne eines gut durchdachten Konzepts mit verschiedenen Phasen, Rückkoppelungen, Reflexionsmöglichkeiten und den dazugehörigen Tools
- ein interdisziplinäres Kernteam an Prozessverantwortlichen und Moderator*innen
- die Einbindung von unterschiedlichen, relevanten Akteuren und Zielgruppen

- das Auffinden von interessanten Schnittstellen zum Thema, die am Alltag (also an der Lebenswelt, Gewohnheit und Erfahrung) der Teilnehmer*innen ansetzen
- demokratische Methoden, die nicht nur Mehrheitsentscheidungen vorsehen (z.B. systemisch Konsensieren) und auf das „Ausräumen von Widerständen“ abzielen
- geeignete Visualisierungen, Vergleiche, Veranschaulichungsmöglichkeiten, spielerische Ansätze, sowie Testmöglichkeiten
- die Sichtbarmachung der erzielten Ergebnisse (bspw. nach dem Prinzip der produktiven Redundanz)
- eine Evaluierung bei Teilnehmenden und Akteuren
- die Überführung in einen weiteren kontinuierlichen Prozess, der vom Kernteam an andere Akteure übergeben wird (Verantwortungsübergang)

5.4. Die Perspektive der EU

Die Etablierung eines fairen und gerechten Regelwerks rund um KI dient gemäß dem Weißpapier der EU dazu, „materielle und immaterielle Schäden von Menschen abzuwenden“, siehe COM(2020). Aus der Perspektive unterschiedlicher Akteurs-Ebenen ist es allerdings schwierig, die Auswirkungen der KI (auch auf verschiedene Grundrechte) zu beurteilen, denn Gefahren und Risiken der KI lassen sich aufgrund rasanter (disruptiver) Entwicklungsschübe nur schwer abschätzen, was Menschen verunsichert und beunruhigt.

Daher spricht das Weißbuch im weiteren Verlauf von der Notwendigkeit der Schaffung eines „einzigartigen Ökosystem für Vertrauen“, wobei die Kommission „ein Konzept, bei dem der Mensch im Mittelpunkt steht“, befürwortet (COM(2020), Seite 3). An anderer Stelle (Seite 19) wird davon gesprochen, dass der Mensch die Ziele bestimmt und programmiert, die ein KI-System erreichen sollen und dass, „indem Europa durch Daten befähigt wird, bessere Entscheidungen zu treffen das Leben seiner Bürgerinnen und Bürger verändert wird“. (Seite 3)

Ein gewünschter europaweiter KI-Regulierungsrahmen im Sinne eines starken Verbraucherschutzes, wird allerdings als komplexes und schwieriges Unterfangen wahrgenommen. Die EU setzt dabei auf einen risikobasierten Ansatz, um der Verhältnismäßigkeit des regulatorischen Eingreifens gerecht zu werden. Sie räumt auch ein, dass Entscheidungen, die von oder mithilfe von KI-Systemen gefällt werden, schwer nachvollziehbar sind und derzeit kaum wirksam angefochten werden können.

Hieraus ergibt sich unserer Überzeugung nach als umso wichtigere Voraussetzung, für die Bürger einen Raum zur Erweiterung des grundsätzlichen technologischen Verständnisses wie auch für Möglichkeiten entsprechender Diskurse und für positive Beispiele der partizipativen Technikgestaltung, aber auch für Ängste, Sorgen und wahrgenommene Bedrohungen und Diskriminierung zu schaffen. Gute Beispiele für faszinierende und anschaulich verständliche Entwicklungen und Anwendungen für KI kommen dabei etwa aus dem Gesundheitsbereich (Augensteuerung von Computertastaturen mit anschließender Sprachassistenz bei gelähmten Personen, Schuhe, die Hindernisse erkennen, Sehbehelfe, die kognitive Barrieren überlisten etc.).

Aus Sicht der High Level Expert Group (HLEG 2020) ergeben sich dabei folgende Kernanforderungen, welche sich auch in den Diskussionen des Expert*innenworkshops widerspiegeln (vgl. Kapitel 4):

- Vorrang menschlichen Handelns und menschlicher Aufsicht
- Technische Robustheit und Sicherheit
- Privatsphäre und Datenqualitätsmanagement
- Transparenz
- Vielfalt, Nichtdiskriminierung und Fairness
- gesellschaftliches und ökologisches Wohlergehen und
- Rechenschaftspflicht

Aus ethischer Sicht wird in von der EU AI High Level Expert Group in (HLEG 2019) der minimalethische Ansatz aus der Bioethik der 1970er Jahre aufgegriffen. Dabei geht es also weniger um das faktisch vorhandene Vertrauen, sondern um das vorhergehende Setzen von Bedingungen der Vertrauenswürdigkeit. Letztlich stellt dies die normative Beschreibung dessen heraus, was wir im Rahmen des Expert*innenworkshops deskriptiv erfasst haben. Die zugrundeliegenden Prinzipien sind dabei auf der einen Seite abstrakt genug, um für möglichst viele Anwendungssituationen zu greifen und auf der anderen Seite noch hinreichend konkret, um situationsspezifische Aussagen zu ermöglichen. Es handelt sich also hierbei um einen Konsens, die pragmatische Kunst des Möglichen in welcher der kleinste gemeinsame Nenner normativer Anforderungen gesucht wird (darum auch als „Minimalethik“ bezeichnet):

- Achtung der menschlichen Autonomie
- Schadensvermeidung
- Fairness
- Transparenz/Erklärbarkeit

In ähnlicher Weise formulierte der Österreichische Rat für Robotik und Künstliche Intelligenz in einem White Paper vom November 2018 einen allgemeinen Werterahmen, der auch für KI gelten soll. Dieser beinhaltet:

- Menschenrechte: Sie sind in nationalen und internationalen Grundrechtskatalogen, Menschenrechtsverträgen und Erklärungen festgelegt und umfassen die Freiheit und Würde des Menschen, wirtschaftliche, kulturelle und soziale Rechte sowie den Schutz der Privatsphäre;
- Gerechtigkeit und Fairness, Inklusion und Solidarität, einschließlich des Schutzes der schwächeren und schutzbedürftigen Bevölkerungsmitglieder;
- Demokratie und Mitbestimmung;
- Grundsatz der Nichtdiskriminierung;
- Persönliche, gesellschaftliche und geteilte Verantwortung;
- Weitere in Österreich und Europa gebräuchliche ethische und gesellschaftliche Werte“ (ACRAI (2018), S. 21)

6. Fazit

Die Besonderheit KI-basierter Entscheidungs- und Optimierungssysteme besteht in der hohen Komplexität von Algorithmen, die einerseits oft schwer verständlich und andererseits im Falle von Biases oder anderer unerwünschter Konsequenzen schwierig korrigierbar sind. Die Faszination der KI beruht dabei u.a. auf dem Staunen und der Bewunderung der technischen Innovationen. Insbesondere der Komfortgewinn und die Verbesserung der Lebensqualität durch verschiedene Assistenzsysteme, die Abläufe optimieren, erleichtern und übernehmen, sind positive Beispiele, die es auch zu veranschaulichen gilt.

Damit aber mögliche Fehler der KI nicht die Schlagzeilen dominieren und KI als Bedrohung wahrgenommen wird, braucht es Interventions- und Verhandlungsräume zum Erreichen von größerer Fairness und mehr Diskurs sowie unterschiedliche Formate für mehr Verständnis von KI, um letzten Endes auch Handlungsspielräume ausloten und schließlich realisieren zu können. Hierzu bedarf es Möglichkeiten der gesellschaftlichen und unternehmensinternen Partizipation, sowohl in der Entwicklung als auch in der Qualitätskontrolle in der Nutzung. Besonders KI-basierte Empfehlungs- und Entscheidungsalgorithmen, welche direkte Auswirkungen auf Menschen und ihre Lebenswelt ausüben, sind kritisch zu hinterfragen, um im Falle von Diskriminierungsverdachtsmomenten weitere Prozessschritte zu erarbeiten und auszuhandeln. Demzufolge empfiehlt es sich, partizipative Prozesse langfristig und kontinuierlich zu etablieren, um das Vertrauen in die Mitgestaltung und einen möglichen Korrekturmechanismus (durch ein breit akzeptiertes Prozessdesign) zu stärken. Dieses „Ökosystem des Vertrauens“ benötigt die aktive Mitgestaltung durch den Menschen.

Von zentraler Bedeutung für die Fähigkeit, in Aushandlungsprozesse und Diskurse über die Fairness von KI-Entscheidungen in einer Gesellschaft einzutreten, ist zunächst die Akzeptanz von demokratischen Werten und Wahrung von Grund und Persönlichkeitsrechten. Ein „faire KI“ braucht darüber hinaus insbesondere das Verständnis und die Befähigung auch der jungen Generation, zukünftig aktiv mitentscheiden zu wollen. Diese Befähigung muss erlernt werden, daher sind vor allem auch „Wissensvermittler*innen“ aller Art wichtige Adressaten für die vorgestellten Projektergebnisse. Dabei ist das Erarbeiten gelingender Diskursformate von entscheidender Bedeutung. Hierzu werden neben bestehenden Austauschformaten, wie sie sich beispielsweise im Rahmen der Initiativen „Digitaler Humanismus“[10] und „Homo Digitalis“[11] in Österreich bereits etabliert haben, insbesondere die in dAIalog.at erarbeiteten partizipativen Formate einen wichtigen Beitrag leisten können.

[10] https://dighum.ec.tuwien.ac.at/workshop2020/
[11] https://homodigitalis.at

Literaturverzeichnis

ACRAI (2018): Die Zukunft Österreichs mit Robotik und Künstlicher Intelligenz positiv gestalten. White Paper, Österreichischer Rat für Robotik und Künstliche Intelligenz, November 2018. Online verfügbar: https://www.acrai.at/wp-content/uploads/2019/04/ACRAI_whitebook_online_2018.pdf.

AI HLEG (2019) Ethics Guidelines for Trustworty Artificial Intelligence. High-Level Expert Group on Artificial Intelligence. 8. April 2019. European Commission, Brüssel [https://ec.europa.eu/futurium/en/ai-alliance-consultation (12. April 2019)].

Allhutter, Doris, et al. (2020): Der AMS-Algorithmus: Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS). ITA-Projektbericht Nr.: 2020-02. Online verfügbar: http://epub.oeaw.ac.at/ita/ita-projektberichte/2020-02.pdf.

Beauchamp, T. L, und J.F. Childress (2001): Principles of Biomedical Ethics. Fifth Edition. Oxford: Oxford University Press.

Bergmann, Matthias, Thomas Jahn, Tobias Knobloch, Wolfgang Krohn, Christian Pohl und Engelbert Schramm (2010): Methoden transdisziplinärer Forschung. Ein Überblick mit Anwendungsbeispielen. Frankfurt a.M. & New York: Campus.

COM(2020): Zur Künstlichen Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen. Weissbuch der Europäischen Kommission, Brüssel, 19.02.2020. Online verfügbar: https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_de.pdf (zuletzt abgerufen: 29.05.2021).

Deutscher Bundestag (2020): Kurzfassung des Abschlussberichts Enquete-Kommission Künstliche Intelligenz. Berlin, 27.10.2020, online verfügbar: https://www.bundestag.de/resource/blob/801584/102b397cc9dec49b5c32069697f3b1e3/Kurzfassung-des-Gesamtberichts-data.pdf (zuletzt abgerufen: 30.05.2021).

Eyert, Florian, und Paola Lopez (2021): KI demokratisieren -Fairness und Transparenz lassen sich nicht durch Technik allein herstellen.Presse-Blog „Die Macht der Daten“, 29.03.2021. Online verfügbar: https://bibliothek.wzb.eu/artikel/2021/f-23705.pdf (zuletzt abgerufen: 05.06.2021).

Funk, Michael (2020): „WHAT IS ROBOT ETHICS? ...AND CAN IT BE STANDARDIZED?“ in: Marco Nørskov, Johanna Seibt & Oliver Santiago Quick (eds.): Culturally Sustainable Social Robotics. Proceedings of Robophilosophy 2020. August 18–21, 2020, Aarhus University and online. (Frontiers in Artificial Intelligence and Applications, 335). Amsterdam a.o.: IOS Press, pp. 469-480.

Funk, Michael (2021a): Rethinking Transdisciplinarity Through Philosophy of Technology. An Epistemological and Methodological Investigation. PhD Thesis, University of Vienna. (forthcoming)

Funk, Michael (2021b): ROBOTER- UND KI-ETHIK. Eine methodische Einführung – Grundlagen der Technikethik Band 1. Wiesbaden: Springer Vieweg (forthcoming).

Funk, Michael (2021c): ANGEWANDTE ETHIK UND TECHNIKBEWERTUNG. Ein methodischer Grundriss – Grundlagen der Technikethik Band 2. Wiesbaden: Springer Vieweg (forthcoming).

Funk, Michael, Christopher Frauenberger und Peter Reichl (2020): „VOM KÜNSTLICHEN LEBEN ZUR LEBENSKUNST – WAS DIE ETHIK DIGITALER BILDUNG VON ÖKOLOGISCHER VERANTWORTUNG LERNEN KANN“. In: Medienimpulse, Bd. 58, Nr. 03 (2020): Nachhaltigkeit, Digitalisierung und Medienpädagogik? (Online first 20 September 2020) [(DOI) 10.21243/mi-03-20-17] [open access].

Handler, Martina (2018): Bürgerbeteiligung in der Praxis. Ein Methodenhandbuch. ÖGUT, Verlag Stiftung Mitarbeit.

Hirsch Hadorn, Gertrude, Holger Hoffmann-Riem, Susette Biber-Klemm, Walter Grossenbacher-Mansuy, Dominique Joye, Christian Pohl, Urs Wiesmann und Elisabeth Zemp (eds.) (2008): Handbook of Transdisciplinary Research. Dordrecht: Springer.

Hubig, Christoph (2007): Die Kunst des Möglichen II. Ethik der Technik als provisorische Moral. Bielefeld: Transcript.

O’Neill, Cathy (2016): Weapons of Math Destruction. UK: Penguin Books.

HLEG (2019): Ethische Leitlinien für eine vertrauenswürdige KI. High Level Expert Group on Artificial Intelligence set up by the European Commission. Online: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (zuletzt abgerufen: 05.06.2021).

HLEG (2020): Assessment List for Trustworthy Artificial Intelligence (ALTAI). High Level Expert Group on Artificial Intelligence set up by the European Commission. Online: https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment (zuletzt abgerufen: 29.05.2021).

Klein, Julie Thompson (2017): “Typologies of Interdisciplinarity. The Boundary Work of Definition.“ In: Frodeman, Robert, Klein, Roberto C.S. Pacheco (eds). 2017: The Oxford Handbook of Interdisciplinarity. Second Edition. Oxford: Oxford University Press, pp. 21-34.

Klein, Julie Thompson (2008): “Integration in der inter- und transdisziplinären Forschung.” In: Bergmann, Matthias & Engelbert Schramm (eds.) (2008): Transdisziplinäre Forschung. Integrative Forschungsprozesse verstehen und bewerten. Frankfurt a.M. & New York: Campus, pp. 93-116.

Mittelstraß, Jürgen (2018): “The Order of Knowledge: From Disciplinarity to Transdisciplinarity and Back.” In: European Review, Vol. 26, No. S2, pp. 68–75. [doi:10.1017/S1062798718000273]

Pohl, Christian, Lorrae van Kerkhoff, Gertrude Hirsch Hadorn & Gabriele Bammer (2008): “Integration.” In: Hirsch Hadorn, Gertrude, Holger Hoffmann-Riem, Susette Biber-Klemm, Walter Grossenbacher-Mansuy, Dominique Joye, Christian Pohl, Urs Wiesmann, Elisabeth Zemp (eds.) 2008: Handbook of Transdisciplinary Research. Dordrecht: Springer, pp. 411-424.

Zweig, Katharina, und T. D. Krafft (2018): Fairness und Qualität algorithmischer Entscheidungen. In R. Mohabbat Kar, B. E. P. Thapa, & P. Parycek (Hrsg.), (Un)berechenbar? Algorithmen und Automatisierung in Staat und Gesellschaft. Berlin: Fraunhofer-Institut für Offene Kommunikationssysteme FOKUS, Kompetenzzentrum Öffentliche IT (ÖFIT), S. 204-227.

Zweig, Katharina (2019): Ein Algorithmus hat kein Taktgefühl. Düsseldorf: Heyne.