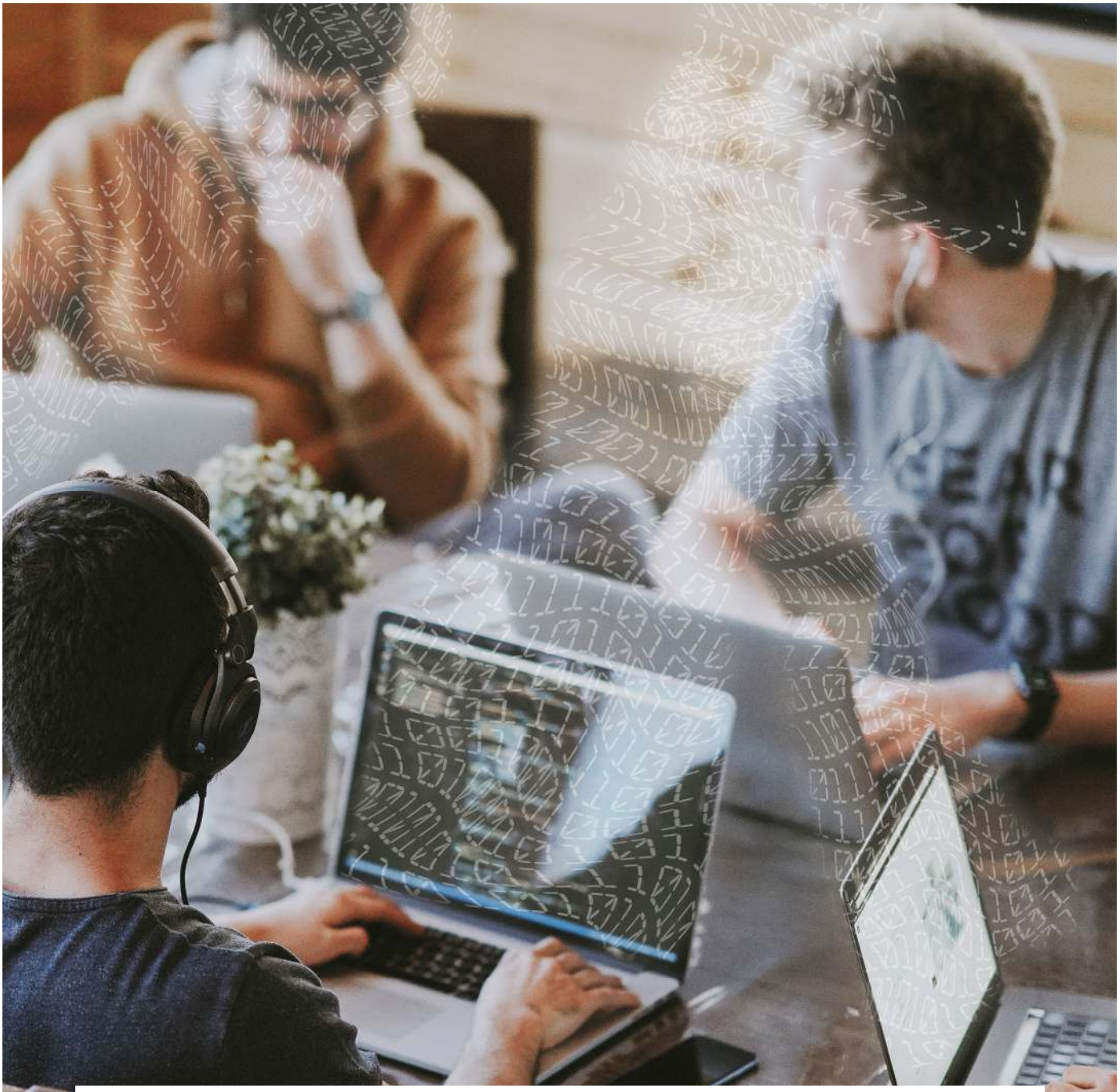


Gütekriterien für algorithmische Prozesse

Eine Stärken- und Schwächenanalyse ausgewählter Forderungskataloge

BertelsmannStiftung

Gütekriterien für algorithmische Prozesse

Eine Stärken- und Schwächenanalyse ausgewählter Forderungskataloge

- Arbeitspapier -

Noëlle Rohde

Impressum
© Juli 2018
Bertelsmann Stiftung
Carl-Bertelsmann-Straße 256
33311 Gütersloh
www.bertelsmann-stiftung.de

Verantwortlich
Ralph Müller-Eiselt

Autorin
Noëlle Rohde

DOI 10.11586/2018027 https://doi.org/10.11586/2018027

Inhalt

1 Vorwort

Was dürfen Algorithmen und was nicht? Welche Qualitätsstandards müssen sie erfüllen? Zu welchen Zwecken dürfen sie eingesetzt werden? Diese Fragen gehen uns alle an. Unsere Gesellschaft muss sich auf Antworten darauf verständigen. Die wichtigste Stellschraube für eine Ethik der Algorithmen sind jedoch die Menschen, die algorithmische Prozesse gestalten. Auftraggeber, Programmierer und Datenwissenschaftler müssen sich ihrer Verantwortung in der Technikgenese und -implementierung bewusst werden und dementsprechend handeln.

Und doch ist es angesichts der beachtlichen Vielfalt der Akteure schwierig, eine bestimme Berufsgruppe als verantwortlich herauszustellen und mit klaren Forderungen zu konfrontieren. Vielversprechender erscheint es da, genau an dem Aspekt anzusetzen, der alle Akteure verbindet – dem algorithmischen Prozess.

Einige hochrangige internationale Arbeitsgruppen, Institute und Vereine haben sich bereits der Herausforderung angenommen, Gütekriterien für Algorithmen zu formulieren. Dies ging augenscheinlich mit beachtlichem Aufwand und hohen Ansprüchen seitens der Beteiligten einher und führte schon zu aussichtsreichen Ergebnissen. Leider ist jedoch zu beobachten, dass diese Gütekriterienkataloge oft lose nebeneinander stehen und neue Prozesse angestoßen werden, ohne aus vorherigen zu lernen. Durch eine solche Arbeitsweise werden nicht nur mögliche Synergieeffekte verschenkt, die Vielzahl verschiedener Ansätze erschwert auch die Herausbildung einer Professionsethik im Sinne zentraler, allenthalben akzeptierter Standards.

Darüber hinaus ist festzustellen, dass die Diskussion über ethische Anforderungen an algorithmische Prozesse und deren Gestaltung überwiegend im internationalen englischsprachigen Raum stattfindet. Die vorliegende Analyse vergleicht drei Gütekriterienkataloge für algorithmische Prozesse: die *Principles for Accountable Algorithms and a Social Impact Statement for Algorithms* der FAT/ML-Konferenz (Fairness, Accountability and Transparency in Machine Learning), die *Asilomar AI Principles* des Future of Life Institute sowie die *Principles for Algorithmic Transparency and Accountability* des ACM (Association for Computing Machinery) US Public Policy Council.

Ziel ist es, auf den deutschen Sprachraum übertragbare Stärken und zu vermeidende Schwächen der Dokumente zu identifizieren. Dazu werden zum einen die Entstehungshintergründe aller drei Vorschläge beleuchtet und zum anderen die konkreten inhaltlichen Forderungen und deren prozedurale Umsetzung bewertet. Die Analyse zeigt auf, welche Kriterien für einen deutschsprachigen Gütekriterienkatalog besonders sinnvoll und welche ungeeignet sind – und wo sich Lücken in den bestehenden Forderungskatalogen finden, die sinnvoll gefüllt werden sollten.

Die drei hier analysierten Gütekriterienvorschläge spiegeln die Vielfalt einschlägiger Ansätze wider: Sie sind das Produkt wissenschaftlicher Konferenzen (FAT/ML) oder direkter Policy-Bemühungen (ACM US Public Policy Council), sie konzentrieren sich auf Einzelaspekte von Algorithmen, wie z. B. Accountability (FAT/ML), oder versuchen das ganze Spektrum im Umgang mit künstlicher Intelligenz abzudecken (Future of Life Institute). Sie haben den Anspruch, unmittelbar handlungsleitend zu sein (FAT/ML, ACM US Public Policy Council) oder befassen sich mit Haltungen und Grundannahmen gegenüber dem neuen Themenfeld (Future of Life Institute).

Wir veröffentlichen diese Analyse als Arbeitspapier, um einen Beitrag zu einem sich schnell entwickelnden Feld zu geben, auf den auch andere aufbauen können. Wir freuen uns über Erweiterungen, Verbesserungen und natürlich auch konstruktive Kritik. Um einen solchen Diskurs zu erleichtern, veröffentlichen wir das Arbeitspapier unter einer freien Lizenz (CC BY-SA 3.0 DE).

Die vorliegende Analyse ist Teil des Projektes „Ethik der Algorithmen“, in dem sich die Bertelsmann Stiftung näher mit den gesellschaftlichen Auswirkungen algorithmischer Entscheidungssysteme beschäftigt. Bislang erschienen sind in der Reihe „Impulse Algorithmenethik“ eine Sammlung internationaler Fallbeispiele (Lischka und Klingel 2017), eine Untersuchung des Wirkungspotenzials algorithmischer Entscheidungsfindung auf Teilhabe (Vieth und Wagner 2017), eine Analyse des Einflusses algorithmischer Prozesse auf den gesellschaftlichen Diskurs (Lischka und Stöcker 2017) sowie ein Papier zu Fehlerquellen und Verantwortlichkeiten in Prozessen

algorithmischer Entscheidungsfindung (Zweig 2018) und ein Gutachten zur Bedeutung der neuen Datenschutzgrundverordnung für automatisierte Entscheidungssysteme (Dreyer und Schulz 2018). Zuletzt wurden eine Umfrage zum Thema „Was Deutschland über Algorithmen weiß und denkt“ (Fischer und Petersen 2018) sowie ein Panorama von Lösungsansätzen, um algorithmische Prozesse in den Dienst der Gesellschaft zu stellen (Krüger und Lischka 2018), veröffentlicht.

Das vorliegende Papier setzt sich näher mit bereits existierenden Vorschlägen für Gütekriterien an algorithmische Systeme auseinander. Darauf aufbauend veröffentlicht die Bertelsmann Stiftung im Sommer 2018 eine weitere Analyse, die der Frage nachgeht, wie Gütekriterienkataloge sinnvoll und effektiv implementiert werden können. Dazu werden verschiedene Professionsethiken aus anderen Feldern in den Blick genommen, ihr Erfolg untersucht und die Übertragbarkeit auf die Arbeit an algorithmischen Systemen überprüft. Zudem erarbeitet die Bertelsmann Stiftung derzeit zusammen mit iRights.Lab in einem größer angelegten Stakeholderprozess einen Vorschlag für einen solchen Gütekriterienkatalog, dessen Veröffentlichung für den Herbst 2018 geplant ist.

Ralph Müller-Eiselt
Senior Expert Taskforce Digitalisierung
Bertelsmann Stiftung

2 Einleitung

Im Zuge des stetig zunehmenden Einflusses algorithmischer Entscheidungsprozesse in gesellschaftlich sensiblen Zusammenhängen wird eine Regulierung automatisierten decision-makings immer wichtiger. Eine aussichtsreiche Möglichkeit, dies zu erreichen, ist es, an die moralische Verantwortung der Beteiligten zu appellieren und ein Verständnis einer Professionsethik zu schaffen. Dabei geht es nicht primär um allgemeine moralische Fragen, sondern vielmehr um solche, die sich unmittelbar aus der eigenen beruflichen Tätigkeit ergeben und für diese kennzeichnend sind.

Neben einem impliziten Ehrenkodex und Selbstverständnis besteht eine weitere Form der Professionsethik in der Orientierung an expliziten Regelwerken, die die Verantwortlichkeiten und Pflichten einer bestimmten Berufsgruppe darlegen. Erfolgreiche Beispiele für dieses Modell sind unter anderem der Hippokratische Eid (Genfer Gelöbnis) für Mediziner sowie der Pressekodex für Journalisten.

Die Profession der an der Gestaltung gesellschaftlich relevanter algorithmischer Prozesse Beteiligten zu definieren, wird vor allem durch die Heterogenität der Akteure erschwert. Die Verantwortung lediglich bei den Programmierern zu verorten, würde zwangsläufig einer Simplifizierung gleichkommen. Eine Vielzahl weiterer Personengruppen, wie Auftraggeber, politische Entscheidungsträger oder die Vertreter von Institutionen sind maßgeblich an der Planung, Auftragserteilung und dem Einsatz algorithmischer Systeme beteiligt (vgl. Zweig 2017) und beeinflussen auch den Handlungsspielraum der Programmierer.

Eine Möglichkeit, der Heterogenität der Verantwortlichen zu begegnen, ist die Annahme eines prozessorientierten Professionsverständnisses, nach dem all jene als verantwortlich angesehen werden, die an der Gestaltung eines algorithmischen Prozesses beteiligt sind. Um diesen Personenkreis zu adressieren, bietet sich folglich vielmehr eine Zusammenstellung von Gütekriterien an den Prozess an als eine Professionsethik, die sich explizit auf das Verhalten der Praktiker bezieht.

Im internationalen Kontext haben sich in den vergangenen Jahren zahlreiche Organisationen und Arbeitsgruppen mit dem Problem der Gütesicherung von algorithmischen Prozessen befasst. Da dieser Diskurs in Deutschland deutlich weniger entwickelt ist, bietet sich die Analyse einiger ausgewählter internationaler Vorschläge an, um daraus Einsichten für die Erstellung eines deutschsprachigen Entwurfes abzuleiten. Für das vorliegende Papier wurden drei aktuelle Dokumente ausgewählt: das Statement on Algorithmic Transparency and Accountability des ACM US Public Policy Council, die Asilomar AI Principles des Future of Life Institute (AI: artificial intelligence) sowie die Principles for Accountable Algorithms und das Social Impact Statement for Algorithms des FAT/ML-Kollektivs.

Im Folgenden werden diese drei Vorschläge für Gütekriterien beschrieben und analysiert. Dies geschieht mit Hinblick auf diese drei Dimensionen:

- Wer sind die Verfasser? Was sind der Entstehungshintergrund und die Ziele? (Beschreibung)
- Was ist der Inhalt? Was wird konkret gefordert? Wer sind die Adressaten? (Beschreibung)
- Was sind die Stärken des Kataloges, wo besteht Verbesserungsbedarf? (Analyse)

Zuletzt werden aus den Ergebnissen übertragbare Stärken und auszugleichende Schwächen für die Erstellung eines deutschsprachigen Gütekriterienkataloges identifiziert.

3 ACM US Public Policy Council (USACM)

Der ACM US Public Policy Council ist ein Teil der Association for Computing Machinery (ACM) und ein unabhängiges Organ, das Themen der Computer- und Informationstechnologie in der öffentlichen Politik in den USA zur Sprache bringt. Der Council versorgt den amerikanischen Kongress, die Verwaltung und die Gerichte mit Experteneinschätzungen sowie Analysen der aktuellen Entwicklungen in ihrem Fachgebiet. Die US-Einheit arbeitet in der Erstellung von Expertisen häufig mit anderen weltweiten ACM Policy Councils zusammen und publiziert bzw. verbreitet die Ergebnisse in Form von Public Policy Statements und Reports. Das Komitee für Computerfragen setzt sich zusammen aus Repräsentanten der Wissenschaft, Managern und Fachleuten aus der Industrie, der Regierung und dem Nonprofitsektor (ACM 2018).

Am 12. Januar 2017 veröffentlichte der ACM US Public Policy Council sein Statement on Algorithmic Transparency and Accountability, welches am 25. Mai desselben Jahres ebenfalls vom ACM Europe Policy Committee unterzeichnet wurde (ACM 2017) und sieben Kriterien an ADM-Prozesse enthält, sodass deren Vorteile verstärkt und die aus ihnen resultierenden potenziellen Schäden minimiert werden.

Hintergrund und Motivation des Papiers ist, wie in der Einleitung von den Autoren angemerkt, die Beobachtung, dass Algorithmen allgegenwärtig sind und die Gefahr bergen, unerwünschte soziale Verzerrungseffekte hervorzubringen. Aus diesem Grund fordern die unterzeichnenden Parteien ADM-Entwickler (Algorithmic Decision-Making: algorithmische Entscheidungsfindung) und politische Entscheidungsträger auf, die vom ACM US Public Policy Concil entwickelten sieben Standards in algorithmischen Systemen zu wahren.

3.1 Beschreibung

Die vom ACM beschriebenen Standards sind: Problembewusstsein (Awareness), Zugang und Behebung (Access and Redress), Verantwortlichkeit (Accountability), Erklärung (Explanation), Datenursprung (Data Provenance), Überprüfbarkeit (Auditability), Validierung und Testing (Validation and Testing).

- Hinter dem Begriff des Problembewusstseins verbirgt sich die Idee, dass sich die Besitzer, Designer, Hersteller, Nutzer und andere Stakeholder darüber im Klaren sein sollten, dass unentdeckte Vorannahmen in gesellschaftlich sensiblen ADM-Prozessen negative Folgen für die Gesellschaft und das Individuum haben können.
- Die Forderungen nach Zugang und Behebung beziehen sich darauf, dass für diejenigen, die nachteilig von einer algorithmisch vorbereiteten Entscheidung betroffen sind, angemessene Mechanismen zur Beschwerde und Überprüfung eingerichtet werden sollten.
- Dass Institutionen für die von ihren Algorithmen getroffenen oder vorbereiteten Entscheidungen verantwortlich gemacht werden sollten, ist die dritte Forderung.
- Der Standard der Erklärbarkeit algorithmischer Mechanismen besteht darin, dass die verantwortlichen Institutionen ermutigt werden sollten, die systeminternen Prozesse und Entscheidungen darzulegen.
- Ähnlich sollten der Ursprung der Daten, mit denen das System trainiert wurde, öffentlich gemacht und die möglicherweise aus dem Datenset resultierenden Vorannahmen überprüft werden. Nur durch den kritischen öffentlichen Blick auf die Daten sei ein maximales Potential für Korrekturen gegeben, wobei es aus Gründen des Datenschutzes zum Teil gerechtfertigt sei, nur einem bestimmten qualifizierten Personenkreis nähere Einblicke in die Daten zu geben.
- Der sechste Qualitätsstandard der Überprüfbarkeit kann laut dem ACM Council nur dann gewährleistet werden, wenn die Modelle, Algorithmen, Daten und Entscheidungen so aufgezeichnet werden, dass sie später nachvollziehbar sind.
- Zuletzt wird vorgeschlagen, dass Institutionen ein sorgsames Test- und Validierungsverfahren für ihre Modelle einsetzen sollten, um frühzeitig mögliche Diskriminierungseffekte zu bemerken. Darüber hinaus sollten die Ergebnisse dieser Tests öffentlich gemacht werden.

Die Adressaten der Gütekriterien werden für jeden einzelnen Bestandteil definiert und reichen von Regulatoren und Institutionen, Designern, Besitzern und Nutzern bis hin zu politischen Entscheidungsträgern.

3.2 Analyse

3.2.1 Stärken

Die Principles for Algorithmic Transparency and Accountability des ACM US Public Policy Council sind in vielerlei Hinsicht ein ernst zu nehmender und gut durchdachter Katalog von Gütekriterien für gesellschaftlich relevante ADM-Prozesse. Ein großer Vorteil der ACM Principles besteht darin, dass sie auf einem komplexen Verständnis der Entwicklung algorithmischer Systeme beruhen. So beschränken sich die Qualitätsstandards nicht nur auf die Programmierung der Software selbst, sondern beziehen sowohl vorgelagerte Schritte, wie das Problembewusstsein und die Datensammlung bzw. das Training des Algorithmus, als auch nachgelagerte Schritte, wie die Anfechtung ungerechter Ergebnisse und das Nachvollziehen von Fehlern, mit ein.

Dieses multidimensionale Verständnis von algorithmischen Prozessen wird auch in der oben erwähnten Vielzahl von Adressaten reflektiert. Anders als dies in der öffentlichen Wahrnehmung häufig der Fall ist, werden nicht nur die Entwickler der Software als Verantwortliche angesehen, sondern auch die Politik, konkrete Auftraggeber, Institutionen und Nutzer. In anderen Worten werden all jene adressiert, die an einem ADM-Prozess gestaltend beteiligt sind. Ein solches prozessorientiertes Professionsverständis ist vor allem deshalb sinnvoll, weil es die komplexe Realität abbildet, statt zu simplifizieren und somit zu riskieren, eine Vielzahl relevanter Akteure zu vernachlässigen.

In ähnlicher Art und Weise beschränken sich die Qualitätsstandards nicht nur auf das Softwaresystem, sondern auch auf die darin verarbeiteten Daten, und decken damit eine häufig vernachlässigte Fehlerquelle ab.

Darüber hinaus zeugen die ACM Principles in einigen Fragen von einer hohen Sensibilität für mögliche Konflikte zwischen den Qualitätskriterien und anderen Anforderungen an algorithmische Systeme. So wird zum Beispiel die Forderung nach Offenlegung der Daten gemildert, um nicht nur algorithmischer Transparenz und Verantwortung, sondern auch dem Datenschutz, Wirtschaftsgeheimnissen oder dem möglichen Systemmissbrauch Rechnung zu tragen.

3.2.2 Schwächen

Nichtsdestotrotz weisen die Principles des ACM Public Policy Councils auch einige Schwächen auf, die in zukünftigen Gütekriterienkatalogen ausgeglichen werden müssen. Zunächst auf rein formeller Ebene folgen die vorgeschlagenen Qualitätsstandards zeitlich nicht logisch aufeinander. Während das Problembewusstsein (1.) sicherlich auf eine der ersten Entwicklungsstufen eines ADM-Systems abzielt, so scheint doch zum Beispiel die Forderung nach Zugang und Behebung (2.) erst nach der Entwicklung und dem Einsatz eines Systems relevant zu sein. Es ist deshalb nicht selbsterklärend, warum dieses Kriterium zum Beispiel dem Verantwortungsbewusstsein von Institutionen vorangestellt wurde. Ähnlich wird das Testen der Software als letzter Punkt aufgeführt, welches jedoch ebenfalls logischerweise vor (2.) anzusiedeln wäre. Auch wenn diese Inkonsistenz keinen weiteren Einfluss auf die Qualität der Kriterien hat, so kann sie doch das Verständnis erschweren. Auch um die Gruppen der Beteiligten anhand der zeitlichen Abfolge ihres Beitrages zu einem bestimmten ADM-Prozess besser abzugrenzen, wäre eine stringente logische Folge wünschenswert.

Des Weiteren wird ein besonders wichtiger Aspekt von algorithmischer Entscheidungsfindung weitestgehend vernachlässigt: der ethisch korrekte Einsatz. Die einzigen Kriterien, die dieses Herzstück einer Ethik der Algorithmen indirekt zur Sprache bringen, sind das Problembewusstsein für „den potenziellen Schaden, den Verzerrungen (biases; Anm. d. Verf.) für Individuen und die Gesellschaft verursachen können“ und die Forderung nach dem Testen von Software, um zu ermessen, ob diese „diskriminatorische Schäden verursacht“ (ACM US Public Policy Council 2017: 2). Der ethisch korrekte Einsatz von algorithmischen Systemen ist jedoch von so großer gesellschaftlicher Bedeutung, dass dieser ein eigenes Qualitätskriterium darstellen müsste. Wie ein solches Kriterium ausgearbeitet werden könnte, wird weiter unten diskutiert.

Eine weitere Schwäche der ACM-Prinzipien ist, dass sie lediglich sehr abstrakte Forderungen enthalten. Dieser Umstand ist vermutlich der Tatsache geschuldet, dass das Feld algorithmischer Entscheidungsfindung außerordentlich groß und divers ist. Es ist eine Herausforderung, konkrete Kriterien zu formulieren, ohne damit Teile des Einsatzgebietes unbeabsichtigt auszuschließen, und dennoch bedeutet ein abstrakter Katalog, dass sich nur mit Mühe anwendbare Handlungsmaximen daraus ableiten lassen.

Sprachlich fällt bei den ACM Principles auf, dass alle Qualitätskriterien im Konjunktiv formuliert sind (z. B. „Regulators should encourage des adoption of mechanisms that …"). Dies weist zugleich auf eines der größten Probleme des Forderungskataloges hin: seine mangelnde Verbindlichkeit. Insbesondere wo wirtschaftliche oder politische Interessen den Qualitätskriterien gegenüberstehen, wie zum Beispiel bei der zeitlich intensiven Validierung und Testung von Software vor der Markteinführung, wird aus den ACM Principles nicht deutlich, warum sich beteiligte Akteure an die vorgeschlagenen Standards halten sollten. Da es sich bei den Kriterien lediglich um Vorschläge handelt, gibt es weder eine Verpflichtung noch Sanktionierungen für die Nichteinhaltung. Obwohl die Einrichtung solcher Instanzen sicherlich die Wirkungsmacht des ACM Public Policy Councils übersteigt, so sind es doch unter anderem diese Elemente, die eine abstrakte Kriteriensammlung von erfolgreichen und etablierten Professionsethiken wie dem Pressekodex unterscheiden können.

Insbesondere die Verantwortung der Politik in der Untermauerung der Autorität der Gütekriterien bleibt unbeantwortet.

Zusammenfassend lässt sich sagen, dass die Principles of Algorithmic Transparency and Accountability des ACM US Public Policy Councils einen zeitgemäßen, differenzierten Versuch darstellen, Standards für ADM-Prozesse zu formulieren. Aufbauend auf einem nuancierten Verständis für den Komplex algorithmischer Entscheidungsfindung wird eine hilfreiche Orientierung für die Verantwortlichen gegeben, die jedoch zum Teil die wirtschaftliche und politische Realität aus den Augen verliert und damit keine Verbindlichkeit schaffen kann.

4 Asilomar Principles des Future of Life Institute

Das Future of Life Institute ist eine von Ehrenamtlichen geführte Forschungseinrichtung, die sich zum Ziel gesetzt hat, existenzielle Risiken zu minimieren. Als existenzielle Risiken werden solche definiert, die große Teile der Menschheit auslöschen können, wie laut den Mitgliedern, auch künstliche Intelligenz (KI) (Future of Life Institute 2018a). Im Future of Life Institute arbeiten Akteure aus der Wissenschaft, dem öffentlichen Leben und der Industrie zusammen (Future of Life Institute 2018b).

Die 23 sogenannten Asilomar AI Principles des Future of Life Institute entstanden auf der Beneficial AI Conference des Jahres 2017 und sind nach dem Veranstaltungsort benannt. Während der Konferenz arbeiteten Vertreter der Wissenschaft und der Industrie sowie Vordenker aus den Bereichen Wirtschaft, Justiz, Ethik und Philosophie zusammen, um Prinzipien für die gesellschaftlich verantwortliche Gestaltung von ADM-Prozessen auszuarbeiten (Future of Life Institute 2017b).

Der Entstehungsprozess begann zunächst mit den Organisatoren der Konferenz, die aus bereits bestehenden Ansätzen zu Gütekriterien algorithmischer Entscheidungsfindung eine Liste der am häufigsten vertretenen Ansichten zum Management künstlicher Intelligenz zusammenstellten. Diese Liste wurde in einer ersten Überarbeitungsstufe allen Teilnehmern der Konferenz im Vorfeld zur Verfügung gestellt und das erhaltene Feedback bzw. Verbesserungsvorschläge eingearbeitet. In der zweiten Stufe wurden auf der Konferenz die so gesammelten Prinzipien erneut in Einzelgruppen diskutiert und überarbeitet. Zuletzt wurden alle Teilnehmer nach ihrer Zustimmung zu jeder einzelnen Version jedes einzelnen Prinzips befragt. Nur diejenigen Prinzipien, die eine Zustimmung von 90 Prozent erreichten, wurden in die finale Liste aufgenommen (Future of Life Institute 2017c).

4.1 Beschreibung

Motivation und Hintergrund der Prinzipien ist laut der Präambel folgende Annahme: „… (Artificial intelligence; Anm. d. Verf.) will offer amazing opportunities to help and empower people in the decades and centuries ahead" (Future of Life Institute 2017a). Es lässt sich somit festhalten, dass die Asilomar AI Principles auf einer optimistischen Vision des gesellschaftlichen Einsatzes von künstlicher Intelligenz (KI) beruhen.

Die Asilomar AI Principles sind in die Bereiche Forschungsfragen, Ethik- und Wertfragen sowie längerfristige Probleme gegliedert. Insgesamt werden 23 Forderungen aufgestellt (aus dem Englischen paraphrasiert).

Der Bereich der Forschungsfragen enthält fünf Forderungen:

1. Das Forschungsziel bei Themen künstlicher Intelligenz soll die Erschaffung dienlicher und nicht ungerichteter Intelligenz sein (Research Goal).
2. Investitionen in künstliche Intelligenz sollten durch die Finanzierung von solcher wissenschaftlicher Forschung begleitet sein, die den zuträglichen Gebrauch künstlicher Intelligenz garantiert. Dazu gehören komplexe Fragen aus den Bereichen Informatik, Wirtschaft, Rechtwesen, Ethik und Sozialwissenschaften (Research Funding).
3. Es sollte einen konstruktiven und gesunden Austausch zwischen KI-Wissenschaftlern und politischen Entscheidungsträgern geben (Science-Policy Link).
4. Eine Kultur der Kooperation, des Vertrauens und der Transparenz sollte zwischen den Wissenschaftlern und Entwicklern von KI gestärkt werden (Research Culture).
5. Die an der Entwicklung von KI beteiligten Teams sollten aktiv kooperieren, um das Umgehen von Sicherheitsstandards zu vermeiden (Race Avoidance).

Dreizehn Forderungen entfallen auf den Bereich der Ethik- und Wertfragen:

6. KI-Systeme sollten während ihrer gesamten Laufzeit sicher und in dieser Hinsicht – wo angemessen und umsetzbar – überprüfbar sein (Safety).
7. Falls ein KI-System Schaden anrichtet, sollte nachvollziehbar sein, warum (Failure Transparency).
8. Jede Beteiligung eines autonomen Systems an justiziellen Entscheidungen sollte eine befriedigende Erklärung beinhalten, die von einer kompetenten menschlichen Autorität überprüfbar ist (Judicial Transparency).
9. Die Designer und Entwickler von fortschrittlichen KI-Systemen sind die Stakeholder für die moralischen Implikationen des Gebrauchs, Missbrauchs und der Handlungen dieser Systeme. Die Stakeholder haben sowohl die Verantwortung als auch die Möglichkeit, diese Implikationen zu gestalten (Responsibility).
10. Hochautonome KI-Systeme sollten so designed werden, dass ihre Ziele und Verhaltensweisen während ihrer gesamten Anwendung mit menschlichen Werten in Einklang gebracht werden können (Value Alignment).
11. KI-Systeme sollten so designed und eingesetzt werden, dass sie mit den Idealen menschlicher Würde, des Rechts, der Freiheit und der kulturellen Vielfalt vereinbar sind (Human Values).
12. Menschen sollten das Recht haben, die Daten, die sie generieren, einzusehen, zu managen und zu kontrollieren in Anbetracht der Macht von KI-Systemen, diese Daten zu analysieren und einzusetzen (Personal Privacy).
13. Die Anwendung von KI auf persönliche Daten darf die reale oder empfundene Freiheit von Personen nicht ungebührlich einschränken (Liberty and Privacy).
14. KI-Technologien sollten so vielen Menschen wie möglich dienen und ihre Teilhabe ermöglichen (Shared Benefit).
15. Der wirtschaftliche Wohlstand, der durch KI gewonnen wird, sollte breit verteilt werden, um der gesamten Menschheit zum Vorteil zu reichen (Shared Prosperity).
16. Menschen sollten wählen können, wie und ob sie Entscheidungen an KI-Systeme delegieren, um von Menschen gewählte Ziele zu erreichen (Human Control).
17. Die Macht, die durch hoch entwickelte KI-Systeme übertragen wird, sollte benutzt werden, um die sozialen und gesellschaftlichen Prozesse, die zu einer gesunden Gesellschaft beitragen, zu respektieren und zu verbessern und nicht, um sie zu unterwandern (Non-subversion).
18. Ein Wettrüsten mit tödlichen autonomen Waffen sollte vermieden werden (AI Arms Race).

Im Bereich der längerfristigen Fragestellungen finden sich fünf abschließende Forderungen:

19. Da kein Konsens besteht, sollten absolute Annahmen über das obere Limit der zukünftigen Fähigkeiten von KI vermieden werden (Capability Caution).
20. Hoch entwickelte KI könnte eine tief greifende Wandlung in der Geschichte des Lebens auf der Erde bedeuten und sollte daher mit angemessener Sorgfalt und mit angemessenem Ressourceneinsatz geplant und gestaltet werden (Importance).
21. Risiken, die durch KI dargestellt werden, insbesondere katastrophische oder existenzielle, müssen Planungs- und Abschwächungsbemühungen unterworfen werden, die dem zu erwartenden Risikoausmaß entsprechen (Risks).
22. KI-Systeme, die sich selbstständig in solcher Art und Weise verbessern oder replizieren, dass ihre Qualität oder Quantität rapide steigen würde, müssen strikten Sicherheits- und Kontrollmaßnahmen unterliegen (Recursive Self-Improvement).
23. Superintelligenz sollte nur im Dienste weithin anerkannter und geteilter ethischer Ideale, sowie zum Wohle der gesamten Menschheit, statt eines einzigen Staates oder einer einzigen Institution entwickelt werden (Common Good).

4.2 Analyse

4.2.1 Stärken

Die 23 Asilomar AI Principles bilden zusammen einen höchst facettenreichen Vorschlagskatalog zur ethischen Gestaltung von ADM-Prozessen. Dessen Komplexität wird zum Beispiel daran deutlich, dass die Kriterien sich nicht nur auf das Wesen algorithmischer Systeme beziehen, sondern auch deren Genese und weitreichenden Folgen berücksichtigen.

Als direkte Folge daraus besteht eine der größten Stärken der Asilomar AI Principles darin, dass sie den Bereich der Forschung in ihre Betrachtung mit einbeziehen. Neben der Festsetzung eines hehren Forschungsziels (das Kreieren von beneficial intelligence) ist besonders die Forschungsförderung berücksichtigt. Diese soll, wie bereits beschrieben, Forschung zum gesellschaftlich zuträglichen Einsatz von künstlicher Intelligenz ermöglichen. In den Principles werden explizit nicht nur die Informationswissenschaften in diesem Kontext benannt, sondern auch Wirtschaftswissenschaften, Rechtswissenschaften, Ethik und Sozialwissenschaften. Auf diese Weise gelingt es den Autoren der Asilomar Principles, eine Vielzahl von Akteuren einzubinden, die üblicherweise als nicht für die Gestaltung von ADM-Prozessen verantwortlich angesehen werden.

Ein weiterer Vorteil der Asilomar Principles besteht darin, dass neben praktischen Forderungen auch Metafragen zur Rolle künstlicher Intelligenz diskutiert werden. Besonders im Abschnitt Ethics and Values werden diese dezidiert beschrieben. Dabei ist hervorzuheben, dass sich die genannten Werte nicht nur auf die öffentlich schon recht breit diskutierten Aspekte wie Transparenz, Privacy, Sicherheit und Verantwortung abzielen. Darüber hinaus werden auch tiefergehende Grundsätze wie der Einsatz von künstlicher Intelligenz zu shared prosperity und shared benefit integriert. Besonders wichtig ist in diesem Zusammenhang, dass die Übereinstimmung von algorithmischen Systemen mit menschlichen Werten (Forderungen zehn und elf) als zwei Gütekriterien mit in den Katalog aufgenommen wurden. Forderung elf kann dabei als inhaltliche Schärfung von Forderung zehn angesehen werden. Ein Alleinstellungsmerkmal der Asilomar Principles ist sicherlich die konkrete Verankerung einiger dieser Wertvorstellungen. Die Forderung, dass algorithmische Systeme sich mit den Idealen menschlicher Würde, Rechten und Freiheiten sowie mit kultureller Diversität in Einklang bringen lässt, ist eine wichtige Erinnerung an die hohe soziale Relevanz künstlicher Intelligenz, die nicht entgegen gesellschaftlicher Werte eingesetzt werden darf.

Insbesondere im Kontext der gesellschaftlichen Ideale lässt sich festhalten, dass die Asilomar Principles an einen Gesetzestext erinnern, der hohe Ziele steckt und zudem die Möglichkeit zur Auslegung bietet. Ein Vorteil dieser Herangehensweise ist, dass die Kriterien so auf eine Vielzahl von Einsätzen künstlicher Intelligenz angewandt werden können, statt nur auf spezifische Felder zuzutreffen. Mögliche Nachteile einer gesetzesartigen Forderung werden weiter unten diskutiert.

Die Asilomar Principles lassen seitens der Autoren ein tief greifendes Verständnis für die Macht algorithmischer Prozesse vermuten. Dies wird nicht nur deutlich, indem Konsequenzen für gesellschaftliche Grundwerte berücksichtigt werden, sondern auch durch die Aufnahme einer eigenen Sektion von Forderungen und Annahmen, die langfristige Probleme adressieren. Besonders der Verweis darauf, dass fortschrittliche künstliche Intelligenz eine profunde Veränderung des Lebens auf der Erde bedeuten kann und dass gleichzeitig katastrophische und existenzielle Risiken durch deren Einsatz entstehen können, ist von großer Bedeutung. Eine solche explizite Aussage über den weitreichenden Einfluss algorithmischer Systeme unterstreicht rekursiv die Unabdingbarkeit des gesamten Forderungskataloges.

Zuletzt lässt sich festhalten, dass die Prinzipien des Future of Life Institute in hohem Maße differenziert sind. Mit 23 separaten Annahmen und Forderungen ist der Katalog mit Abstand der umfangreichste der analysierten Vorschläge. So wird beispielsweise anstelle des öffentlich häufig vernommenen Rufes nach der Transparenz algorithmischer Systeme zwischen Failure Transparency und Judicial Transparency unterschieden (zur Erklärung s. o.). Eine solche Herangehensweise erlaubt die genauere Parametrisierung ausgewählter Ideale und ist mit verantwortlich für den Gesetzescharakter der Asilomar Principles.

4.2.2 Schwächen

Einer der gravierendsten Nachteile der Asilomar AI Principles ist die Tatsache, dass die Adressaten der Gütekriterien nicht ausreichend kenntlich gemacht werden. Aus einigen Forderungen wie zum Beispiel dem Science-Policy Link (Nr. 3) lässt sich erkennen an wen sich das jeweilige Kriterium richtet – in diesem Fall an Forscher und politische Entscheidungsträger. Dies ist jedoch nur selten der Fall. Insbesondere die Forderungen bezüglich Ethics and Values sowie Longer-Term Issues enthalten keinen Verweis auf mögliche Adressaten bzw. die Verantwortlichen in der Einhaltung der Gütekriterien. Lediglich an einer Stelle wird in diesem Zusammenhang von Designern und Entwicklern gesprochen (Nr. 9), die als Stakeholder für die moralischen Implikationen von künstlicher Intelligenz identifiziert werden.

Angesichts der großen Diversität des Einsatzes algorithmischer Entscheidungssysteme ist eine solche Limitierung auf Informatiker als Alleinverantwortliche nicht tragbar. Vielmehr muss eine Vielzahl von Akteuren in den Blick genommen werden, die direkt oder indirekt an der Gestaltung gesellschaftlich relevanter algorithmischer Prozesse beteiligt sind. Forscher und politische Entscheidungsträger sollten zum Beispiel nicht nur im Bereich von Forschungsfragen adressiert werden, sondern durchaus auch in Fragen der ethischen Kriterien für künstliche Intelligenz. Darüber hinaus werden weitere wichtige Stakeholder, wie öffentliche Institutionen, Auftraggeber und Anwender, gar nicht als Adressaten erwähnt. Durch die mangelnde Sensibilität für die Diversität der Verantwortlichen und die daraus resultierende unzureichende direkte Ansprache besteht die Gefahr, dass die Asilomar AI Principles von vielen der relevanten Akteure als nicht für sie relevant wahrgenommen werden.

In ähnlicher Weise wird auch der öffentliche Diskurs als ein mögliches Format zur ethischen Gestaltung künstlicher Intelligenz vollkommen übersehen.

Wie bereits erwähnt, sind die Asilomar Principles in ihren Formulierungen recht offen gehalten. Obwohl dies eine große Flexibilität in der Anwendung erlaubt, die angesichts der Vielschichtigkeit algorithmischer Prozesse sicherlich hilfreich ist, besteht oft die Gefahr allzu genereller und vager Kriterien. So wird beispielsweise ein Austausch zwischen AI-Forschern und politischen Entscheidungsträgern gefordert, der constructive and healthy ist. Ohne jegliche Parametrisierung kann ein solches Kriterium nur schwerlich als ethische Richtlinie dienen, nicht zuletzt weil die Definition dessen, was als konstruktiv und gesund angesehen wird, in hohem Maße von der Perspektive der Beteiligten abhängt.

In ähnlicher Weise gelingt es den Autoren der Asilomar Principles häufig nicht, Angaben zur praktischen Umsetzung ihrer Forderungen zu machen. So wird zum Beispiel gefordert, Entwicklerteams für künstliche Intelligenz sollten aktiv kooperieren, um das wettbewerbsgetriebene Unterwandern von Sicherheitsstandards zu vermeiden. Wo sich jedoch die wirtschaftlichen Interessen mit ethischen Forderungen gegenüberstehen, ist unklar, wie eine solche Forderung zur Handlungsmaxime werden kann. An dieser Stelle könnten die Asilomar Principles von der Erwähnung konkreter politischer Maßnahmen profitieren.

Kurz gefasst: bei den Asilomar AI Principles des Future of Life Institute handelt es sich um einen hochdifferenzierten Katalog von Gütekriterien, der eine große Sensibilität für die wichtigsten Grundsatz- und Praxisfragen im Bereich ADM erkennen lässt. Die Kriterien gehen weit über die Anforderungen an ein algorithmisches System hinaus und beziehen indirekt auch Akteure (wie z. B. Sozialwissenschaftler) ein, die in der gesellschaftlichen Gestaltung algorithmischer Entscheidungsfindung häufig vergessen werden.

Auf den ersten Blick wirken einige der Forderungen sehr idealistisch und etwas realitätsfern, insbesondere da keine Tools oder Praktiken zur Umsetzung vorgeschlagen werden. Allerdings lässt sich auch argumentieren, dass ein solcher Gütekriterienkatalog wie ein Gesetzestext zu verstehen ist, der die hohen gesellschaftlichen Ideale fixiert und dennoch Auslegungsspielraum bietet, um Einzelfällen im komplexen Feld der algorithmischen Entscheidungsfindung gerecht zu werden.

5 FAT/ML Principles for Accountable Algorithms and a Social Impact Statement for Algorithms

FAT/ML steht für Fairness, Accountability and Transparency in Machine Learning und bezeichnet eine jährliche Konferenz, bei der relevanten Wissenschaftlern und Praktikern die Möglichkeit gegeben wird, die neuen gesellschaftlichen Herausforderungen im Feld selbstlernender Maschinen zu adressieren und zu diskutieren. Die Konferenz wird von einem dauerhaften Komitee aus Wissenschaftlern mehrerer Universitäten sowie von Google und Microsoft Research organisiert und fand 2017 bereits zum vierten Mal statt (FAT/ML 2018).

Das Paper Principles for Accountable Algorithms and a Social Impact Statement for Algorithms wird als Hintergrunddokument auf der FAT/ML-Website gelistet und entstand im Rahmen des sogenannten Dagstuhl-Seminars Data, Responsibly vom 17. bis 22. Juni 2016. Die 39 Teilnehmer des Seminars (darunter zehn Frauen) setzten sich zusammen aus Wissenschaftlern verschiedener internationaler Universitäten und Institute – insbesondere aus dem nord- und südamerikanischen und europäischen Raum – sowie Vertretern von Google und Microsoft. Die Teilnehmer repräsentierten dabei verschiedene Forschungsfelder der Computerwissenschaften, wie zum Beispiel data management, data minging, Sicherheit und Datenschutz und Computernetzwerke. Auch Sozialwissenschaftler, Datenjournalisten und Think-Tank-Vertreter waren Teil des Seminars. Das Paper zur Algorithmic Accountability wurde von dreizehn Teilnehmern unterzeichnet (FAT/ML 2016).

Hintergrund des Dagstuhl-Seminars waren die Anerkennung des weitreichenden gesellschaftlichen Einflusses von Big Data und der in diesem Zusammenhang eingesetzten Technologien sowie die Forderung, diese verantwortungsvoll und in Übereinstimmung mit den ethischen und moralischen Normen der Gesellschaft anzuwenden. Um diesem Ideal näher zu kommen, besteht laut den Organisatoren des Seminars ein „an urgent need to define a broad and coordinated computer science research agenda in this area“ (ebd.). Einen ersten Entwurf dieser Agenda auszuarbeiten, war das erklärte Ziel des Dagstuhl-Seminars – das Ergebnis sind die Principles for Accountable Algorithms and a Social Impact Statement for Algorithms.

5.1 Beschreibung

Gemäß dem Namen gliedert sich das Dokument der FAT/ML in zwei Sektionen: die Principles for Accountable Algorithms und ein Social Impact Statement for Algorithms.

Den Prinzipien sind eine Präambel und eine Prämisse vorangestellt. Erstere enthält die Beobachtung, dass die Verbreitung von automatisierten Entscheidungsprozessen stetig zunimmt und algorithmisch informierte Entscheidungen nun einen beträchtlichen gesellschaftlichen Einfluss ausüben können. Ziel der Principles for Accountable Algorithms sei es, Entwicklern und Produktdesignern dabei zu helfen, algorithmische Systeme so zu designen und zu implementieren, dass sie öffentlich nachvollziehbar sind (publicly accountable). Der Begriff der Accountability beziehe sich in diesem Kontext auf die Verpflichtung, algorithmische Entscheidungsprozesse zu dokumentieren, zu erklären oder zu rechtfertigen, sowie negative soziale Konsequenzen abzumildern.

Die Prämisse hingegen besteht im Kern in der These, dass Algorithmen und die von ihnen verwendeten Daten von Menschen geschaffen werden. Das hieße, für jede automatisierte Entscheidung sei letztlich ein Mensch verantwortlich und die Ausrede „der Algorithmus war’s“ sei nicht gültig.

Aufbauend auf der Prämisse werden fünf Prinzipien für Accountable Algorithms vorgestellt:

- *Verantwortlichkeit (Responsibility)*
  Es müssen äußerlich sichtbare Möglichkeiten geschaffen werden, um für das Individuum oder die Gesellschaft nachteilige Effekte revidieren zu lassen. Zudem muss eine interne Rolle für diejenige Person festgeschrieben werden, die verantwortlich für die schnelle Behebung solcher Probleme ist.

- *Erklärbarkeit (Explainability)*
  Es muss sichergestellt werden, dass algorithmische Entscheidungen sowie jegliche Datengrundlage den Endusern und Stakeholdern in nicht technischer Sprache erklärt werden können.
- *Genauigkeit (Accuracy)*
  Fehlerquellen und Unsicherheiten des Algorithmus und seiner Datengrundlage müssen identifiziert, protokoliert und artikuliert werden, sodass erwartete und Worst-Case-Szenarios verstanden und Gegenmaßnahmen entwickelt werden können.
- *Nachvollziehbarkeit (Auditability)*
  Interessierten Dritten muss die Möglichkeit gegeben werden, das Verhalten des Algorithmus zu hinterfragen, zu verstehen und zu überprüfen. Dies muss geschehen durch die Freigabe von Informationen, die die Überwachung, die Überprüfung oder die Kritik am System ermöglichen. Solche Informationen sind zum Beispiel detaillierte Protokolle, technisch angemessene APIs (Application Programming Interface: Anwendungsprogrammierschnittstelle) und permissive Nutzungsbedingungen.
- *Fairness*
  Es muss sichergestellt werden, dass algorithmische Entscheidungen keine diskriminierenden oder ungerechten Auswirkungen haben, wenn verschiedene Demographien (z. B. Ethnie, Geschlecht usw.) verglichen werden.

Die Prinzipien schließen mit dem Hinweis ab, dass diese bewusst unterspezifiziert seien, um eine breite Anwendung in verschiedensten Kontexten zu ermöglichen. Zwei wichtige Prinzipien seien bewusst ausgelassen worden, da diese bereits an anderer Stelle ausführlich behandelt worden seien. Es handelt sich dabei um Datenschutz (mit Verweis auf die OECD Privacy Principles) und die Auswirkungen menschlicher Experimente (mit Verweis auf den Belmont Report).

Die zweite Sektion des FAT/ML-Dokumentes besteht in einer Sammlung von Fragen und Schritten für die Anfertigung eines Social Impact Statement for Algorithms, das die Einhaltung der oben genannten Prinzipien gewährleisten soll. Die Erstellung eines solchen Statements durch die Algorithmenentwickler solle drei Mal während eines Design- und Entwicklungsprozesses erfolgen: während der Designphase sowie vor der Markteinführung (pre-launch) und danach (post-launch). Nach der Markteinführung des algorithmischen Systems solle zudem das Social Impact Statement veröffentlicht werden, um der Öffentlichkeit die Einschätzung der sozialen Auswirkungen des Systems zu erlauben.

Der Aufbau des Social Impact Statements orientiert sich erneut an den fünf oben beschriebenen Prinzipien – für jedes von ihnen werden mehrere Leitfragen spezifiziert sowie erste Schritte zu deren Erreichung angegeben. So wird zum Beispiel im ersten Prinzip der Verantwortlichkeit gefragt, welche Person für Schäden an Nutzern verantwortlich ist und welche Berichts- bzw. Regressmöglichkeiten es gibt. Um diese Fragen zu klären, sind als konkrete Schritte unter anderem die Benennung eines Verantwortlichen für die sozialen Auswirkungen des Algorithmus, die Bereitstellung von Kontaktinformationen für Beschwerden sowie die Entwicklung eines Plans für die Reaktion auf ungewollte Nebeneffekte des Systems benannt.

Weitere Maßnahmen aus dem Social Impact Statement sind beispielsweise folgende:

- *Erklärbarkeit:* einen Plan erstellen, in dem der Entscheidungsprozess den Nutzern erklärt wird; Betroffenen den Zugang zu ihren Daten ermöglichen und ihnen die Möglichkeit geben, diese bei Bedarf zu ändern; bei Machine-Learning-Modellen die Trainingsdaten beschreiben; die Veränderung und Bereinigung der genutzten Daten erklären.
- *Genauigkeit:* das Fehler- und Schadenspotenzial des Systems bewerten; eine Sensitivitätsanalyse durchführen, die Aufschluss darüber gibt, wie die Unsicherheit des Outputs mit der Unsicherheit des Inputs korreliert; einen Prozess entwickeln, mit dem Menschen Fehler in den Input- und Trainingsdaten sowie in den Output-Entscheidungen revidieren können; einen Validitätstest durchführen, um die gesamte Datenfehlerquote am Beispiel eines zufälligen Sets zu veröffentlichen.
- *Nachvollziehbarkeit:* eine API dokumentieren und bereitstellen, die es Dritten erlaubt, das algorithmische System zu hinterfragen und zu bewerten; sicherstellen, dass die Nutzungsbedingungen es der wissenschaftlichen Gemeinschaft erlauben, automatisierte öffentliche Audits durchzuführen; einen Plan zur Kommunikation mit Dritten erstellen, die an der Überprüfung des algorithmischen Systems interessiert sind.
- *Fairness:* den Kontakt zu Experten suchen, die mit dem subtilen sozialen Kontext vertraut sind, in dem das System (sich) formiert; überprüfen ob die individuelle Zugehörigkeit zu bestimmten Kategorien (z. B. Ethnie, Geschlecht, Genderidentität, Religion, sozioökonomischer Status etc.) sich nachtteilig auf die durch den Algorithmus erreichten Ergebnisse auswirkt.

5.2 Analyse

5.2.1 Stärken

Da die FAT/ML Principles sich ausschließlich auf den Aspekt der Accountability konzentrieren, besteht eine klare Stärke des Dokumentes darin, dieses Kriterium detaillierter zu beleuchten, als dies in anderen Gütekriterienkatalogen der Fall ist. In den Forderungen des ACM US Public Policy Council ist die Rechenschaft (accountability) als ein einzelner Aspekt genannt, der – wie auch in den Asilomar Principles – mit Verantwortlichkeit (responsibility) gleichgesetzt zu sein scheint. Dass ein wesentlich differenzierterer Begriff von Accountability zugrunde gelegt wird, der sich aus weiteren Kriterien zusammensetzt, ist eine der besonderen Leistungen der FAT/ML Principles. Die fünf konstitutiven Kriterien für Accountability zielen in ihrer Gänze zudem sowohl auf technische (Verantwortlichkeit, Erklärbarkeit, Genauigkeit, Nachvollziehbarkeit) als auch auf moralische Angemessenheit (Fairness) ab.

Ein weiterer Vorteil des Dokumentes besteht in der Grundannahme, expliziert in der Prämisse, dass die Verantwortung für die Entscheidung algorithmischer Systeme immer bei Menschen verortet ist. Besonders im Zeitalter von Medienberichten über die Machtlosigkeit menschlicher Akteure gegenüber Maschinen und der von Unternehmen und Institutionen allzu oft genutzten Entschuldigung „der Algorithmus ist schuld“, ist dies eine unabdingbare Grundlage für eine Professionsethik.

Die wichtigste Stärke des FAT/ML-Kataloges besteht in der praktischen Umsetzbarkeit der vorgeschlagenen Gütekriterien. Das Vorhandensein eines konkreten Entwurfes für ein Social Impact Statement unterscheidet dieses Dokument deutlich von den zuvor analysierten Vorschlägen. Nicht nur wird die Form eines solchen Statements beschrieben, sondern es werden auch konkrete zu beantwortende Fragen und wichtige Schritte angegeben. Auf diese Weise wird das Erstellen eines Social Impact Statements deutlich erleichtert und es gelingt mit dem FAT/ML-Dokument, die häufig rein abstrakte Natur anderer Gütekriterienkataloge zu überwinden.

Zudem verlangen die Autoren die Veröffentlichung des Social Impact Statements seitens der Verantwortlichen, um einen öffentlichen Diskurs über ein algorithmisches System zu ermöglichen. Wie bereits oben erwähnt, sollte die informierte Öffentlichkeit als Akteur in der Gestaltung algorithmischer Systeme nicht unterschätzt werden, da sie das Forum für eine Diskussion bietet, die über wirtschaftliche Interessen hinausgeht. Dennoch kann es häufig sinnvoll sein, keine vollständigen Einblicke in die Funktionsweise und Datengrundlage einer Software zu geben, um Industriespionage, Datenrechtsverletzungen und Manipulationen des Systems zu vermeiden. Ein Social Impact

Statement erlaubt es der Öffentlichkeit, eine Einschätzung der gesellschaftlichen Relevanz und des Gefahrenpotenzials zu erlangen, ohne dass sensible Daten in falsche Hände geraten können.

Zuletzt lässt sich festhalten, dass im FAT/ML-Katalog nicht nur Kriterien an den Programmierungsvorgang gerichtet werden, sondern sinnvollerweise auch nachgelagerte Schritte, wie zum Beispiel die Anfechtung falscher Entscheidungen, Berücksichtigung finden.

5.2.2 Schwächen

Anschließend an die oben zuletzt genannte Beobachtung ist leider zu bemängeln, dass im FAT/ML-Dokument zwar auf nachgelagerte, aber nicht explizit auf vorgelagerte Schritte in der Gestaltung algorithmischer Prozesse eingegangen wird. So fehlt beispielsweise der Rekurs auf die Forschung zu algorithmischen Prozessen und deren moralische Integrität gänzlich.

Wie ebenfalls bereits oben erwähnt, befassen sich die Autoren des FAT/ML-Kataloges explizit mit dem Begriff der Accountability, was zu einem differenzierten Bild dieses Aspektes beiträgt. Dennoch werden weitere wichtige Gütekriterien für Algorithmen nicht in den Blick genommen, wie beispielsweise die Sicherheit des Systems oder der Datenschutz. Insbesondere die moralischen Anforderungen an das Gestalten algorithmischer Prozesse sind nicht hinreichend berücksichtigt. Aus den fünf Gütekriterien sind, wie bereits erwähnt, fünf technischer Natur und nur der Aspekt der Fairness zielt auf die ethische Dimension der Gestaltung.

Dass ein Algorithmus fair entscheidet (bzw. solche Entscheidungen vorbereitet), ist sicherlich eine wichtige Anforderung in gesellschaftlich sensiblen Fragen. Dennoch ist Fairness lediglich ein komparatives Konzept – es besagt, dass jede Entscheidung auf Grundlage desselben Sets von Parametern getroffen werden muss. Einzelne Individuen dürfen beispielsweise nicht aufgrund ihrer Hautfarbe oder ihres Geschlechts benachteiligt werden, wenn diese Kriterien bei der Auswahl anderer nicht zu einer Abwertung führen. Es ist allerdings festzuhalten, dass nicht jede unmoralische Entscheidung automatisch unfair sein muss – prinzipiell kann sie für jedes Individuum gleichermaßen gegen moralische Regeln verstoßen. In dem FAT/ML-Katalog ist diese Dimension leider nicht berücksichtigt und so fehlt beispielweise der Bezug auf die Menschenrechte, die moralischen Werte einer Gesellschaft oder das Grundgesetz gänzlich.

Zuletzt lässt sich festhalten, dass das FAT/ML-Statement auf einem recht engen Verständnis vom Kreis der Verantwortlichen für algorithmische Systeme beruht. Die Präambel richtet sich explizit nur an Developer und Produktdesigner und das Social Impact Statement soll lediglich durch „Algorithmenbauer“ verfasst werden. In Anlehnung an Zweig (2018) lässt sich eine Vielzahl weiterer Akteure identifizieren, die an der gesellschaftlichen Gestaltung von Softwaresystemen beteiligt ist. Auftraggeber, politische Entscheidungsträger, öffentliche Institutionen und Implementierer sind nur einige der Gruppen, die leider durch die FAT/ML Principles nicht angesprochen werden.

Zusammenfassend lässt sich festhalten, dass es sich bei den FAT/ML Principles um einen Gütekriterienkatalog handelt, der den Aspekt der Accountability in differenzierter Weise parametrisiert und detailliert in seiner Anwendung auf algorithmische Systeme diskutiert. Durch diesen engen Fokus werden jedoch einige weitere wichtige Kriterien vernachlässigt und mehr Sensibilität für ethische (und nicht nur technische) Probleme in der gesellschaftlichen Gestaltung von Algorithmen wäre wünschenswert.

Die herausragende Bedeutung des FAT/ML-Dokumentes liegt in der Skizzierung eines Social Impact Statements. Dieses liefert konkrete Schritte zur öffentlichen Dokumentation und Abschätzung der gesellschaftlichen Bedeutung eines algorithmischen Systems. Durch die Sammlung von Handlungsvorschriften wird somit das Problem der zu generellen Formulierung von abstrakten Kriterien umgangen, das viele andere Gütekriterienkataloge kennzeichnet.

6 Fazit und Ableitungen

Im Folgenden werden aus der Analyse der oben dargestellten Dokumente Ableitungen für die Erstellung eines deutschsprachigen Gütekriterienkataloges vorgenommen. Dabei steht die Beantwortung der folgenden Fragen im Vordergrund:

a) Übertragbare Stärken: Was sind sinnvoll zu übernehmende Annahmen, Inhalte und Formate eines Gütekriterienkataloges für die Gestaltung gesellschaftlich relevanter algorithmischer Prozesse? Dabei wird auch auf dasjenige Dokument verwiesen, in dem der jeweilige Aspekt beispielhaft eingearbeitet wurde. Die Stärken werden in folgenden Kategorien betrachtet:
   - Formale Merkmale
   - Inhaltliche Merkmale (insbesondere technische und moralische Kriterien an algorithmische Prozesse)
   - Angaben zur Implementierung

b) Auszugleichende Schwächen: Welche wichtigen Aspekte wurden in den analysierten Dokumenten nicht berücksichtigt?

6.1 Übertragbare Stärken

Um das Erstellen eines Gütekriterienkataloges formal zu rechtfertigen, ist auf der Ebene der Vorannahmen eine wichtige Einsicht, dass die Verantwortung für algorithmische Prozesse ausschließlich bei menschlichen Akteuren zu verorten ist. Das Voranstellen dieser Annahme in einer Präambel begründet die Motivation für einen Gütekriterienkatalog, der sich an menschliche Gestalter algorithmischer Prozesse richtet. Darüber hinaus wird mit dieser Prämisse ein Kontrapunkt zum in den Medien weit verbreiteten Machtlosigkeitsszenario gegenüber Maschinen gesetzt, ebenso wie zu einer möglichen Verweigerung der Verantwortung seitens der an der Gestaltung Beteiligten. Eine gut geeignete Prämisse ist in den FAT/ML Principles zu finden.

Als eine weitere Prämisse ist es wichtig, die gesellschaftliche Bedeutung von algorithmischen Prozessen zu skizzieren oder zumindest als Grundannahme zu formulieren. Indem die Teilhaberelevanz von Softwaresystemen unterstrichen wird, lässt sich die Bedeutung eines Kodexes für deren Gestaltung rechtfertigen und Akteure können zur Einhaltung der Kriterien motiviert werden. In allen drei Dokumenten ist diese Prämisse vorhanden, allerdings unterscheiden sie sich in ihrer Herangehensweise teilweise sehr stark. Das Papier des ACM US Public Policy Council enthält lediglich eine kurze Beschreibung der Verbreitung algorithmischer Systeme, ebenso wie die FAT/ML Principles, die zudem auf deren „potential for significant societal impact“ (FAT/ML 2018) verweisen. Anders ist die Herangehensweise in den Asilomar Principles des Future of Life Institute. Da der inhaltliche Fokus des Dokumentes auf künstlich intelligenten Systemen (AI, artificial intelligence) liegt, wird hier von einem gesellschaftlich weitreichenden und potenziell verheerenden Einfluss von AI gesprochen. Zudem wird darauf verwiesen, dass keine starken Annahmen über das obere Limit der Fähigkeiten künstlicher Intelligenz getroffen werden sollten, da das gesicherte Wissen darüber zu gering sei. Diese Aussagen finden sich in den Asilomar AI Principles, im Gegensatz zu den zwei weiteren Dokumenten, nicht in der Präambel, sondern als eigenständige Prinzipien in der Sektion Longer-Term Issues. In der Einleitung findet sich jedoch folgende optimistische Annahme: „… guided by the following principles [AI; Anm. d. Verf.] will offer amazing opportunities to help and empower people in the decades and centuries ahead“. Ob eine positive, neutrale oder negative Grundhaltung gegenüber der gesellschaftlichen Bedeutung algorithmischer Prozesse eingenommen wird, ist selbstverständlich eine inhaltliche Frage – dennoch sollte sie, vorzugsweise in einem Vorwort, dem eigentlichen Gütekriterienkatalog zur besseren Nachvollziehbarkeit der Motivation der Autoren vorangestellt werden.

In Bezug auf die inhaltlichen Forderungen lässt sich ebenfalls eine Reihe von Stärken aus den analysierten Dokumenten ableiten.

Zunächst einmal ist es wichtig, der Vielschichtigkeit des algorithmischen Gestaltungsprozesses Rechnung zu tragen und nicht nur den Schritt der Programmierung, sondern auch vor- und nachgelagerte Schritte in den Gütekriterien zu erfassen. In den oben analysierten Dokumenten werden so beispielsweise die Zielsetzung des Einsatzes, die Datensammlung und das Training eines künstlich intelligenten Systems berücksichtigt, wie auch die Möglichkeit, Fehler nachzuvollziehen und anzufechten. Eng damit verknüpft ist auch die Einsicht, dass Fehlerquellen nicht zwangsläufig im System verortet sein müssen, sondern auch durch die Datengrundlage verursacht werden können. Die sich daraus ergebenden Anforderungen an die Qualität der Daten werden besonders im Papier des ACM US Public Policy Council und in den Prinzipien der FAT/ML behandelt.

Wird die Diversität des Gestaltungsprozesses algorithmischer Systeme in dem Gütekriterienkatalog anerkannt, so ergibt sich daraus als direkte Folge ein größerer Adressatenkreis des Dokumentes. Nur wenn nicht lediglich Programmierer, sondern auch Politiker, Auftraggeber, Institutionen und teilweise sogar Anwender angesprochen werden, kann die Gesamtheit des Gestaltungsprozesses erfasst werden. Besonders wird dies in den Asilomar AI Principles geleistet. Für die Erstellung eines deutschsprachigen Gütekriterienkataloges ist es von Vorteil, möglichst viele Akteure aus verschiedenen Bereichen in den Geneseprozess einzubeziehen, um zu einem repräsentativen Dokument zu gelangen.

Zwei Gruppen von Akteuren bedürfen in diesem Zusammenhang gesonderter Aufmerksamkeit: die Forschung und die Öffentlichkeit. Forscher im Bereich algorithmischer Entscheidungsfindung und künstlicher Intelligenz sind ultimativ dafür verantwortlich, was sich in Zukunft als technisch machbar erweist. Lediglich in den Asilomar AI Principles wird die Forschung explizit in den Blick genommen und zu der Entwicklung von „beneficial intelligence“ (Future of Life Institute 2017a) verpflichtet. Diese Sensibilität für die Schlüsselrolle der Forschung ist beispielhaft und sollte für die Entwicklung eines deutschsprachigen Gütekriterienkataloges maßgebend sein. Auch die Forschung in Bereichen außerhalb der Informatik kann maßgeblich zur Gestaltung algorithmischer Prozesse beitragen und es ist deshalb als höchst sinnvoll zu bewerten, dass in den Asilomar AI Principles die Forschungsförderung von relevanten Projekten in den Sozialwissenschaften, der Volkswirtschaft, Ethik und den Rechtwissenschaften gefordert wird.

Die Öffentlichkeit als weiterer wichtiger Akteur wird lediglich in den Prinzipien der FAT/ML und des ACM Public Policy Council in den Blick genommen. Die Förderung eines öffentlichen Diskurses über die Teilhaberelevanz bzw. die Angemessenheit eines algorithmischen Systems ist vor allem deshalb von großer Bedeutung, weil sich so auch (voraussichtlichen) Betroffenen eine Diskussionsplattform bietet. Darüber hinaus lässt sich durch die Veröffentlichung von Social Impact Statements (gefordert in den FAT/ML Principles) vermeiden, dass zwischen konkurrierenden Anbietern ein Wettrennen auf Kosten der Sicherheit (angemerkt in den Asilomar AI Principles) stattfindet und das Gemeinwohl als wichtiges Ziel algorithmischer Prozesse aus den Augen verloren werden könnte. Ein weiterer Vorteil des Einbindens der Öffentlichkeit in die Gestaltung algorithmischer Prozesse besteht darin, dass sich die Zahl derjenigen die ein System auf Fehler überprüfen, enorm erhöht, wie vom ACM US Public Policy Council angemerkt wird. Allerdings ist dabei wichtig, dass Fehler nicht erst im Produktiveinsatz und damit mit weitreichenden Folgen für Betroffene zum Tragen kommen. Der gesellschaftliche Einsatz ist zu hoch für ein solches *Trial and Error* Vorgehen (vlg. Robo-Debt Skandal in Australien, in Rohde 2017).

Aus der Berücksichtigung einer Vielzahl von Akteuren in der Gestaltung algorithmischer Prozesse ergibt sich zwangsläufig die Gefahr von Interessenkonflikten. Diese nicht zugunsten der Einfachheit eines Gütekriterienkataloges zu verschweigen, ist unabdingbar, um die Glaubwürdigkeit des Dokumentes zu wahren. Selbstverständlich lassen sich die möglichen Szenarien nicht in Gänze beschreiben, weshalb sich ein Verweis auf die Problematik nach dem Beispiel des ACM US Public Policy Council anbietet. Dort wird in der Einleitung erwähnt und anhand von kurzen Beispielen illustriert, dass technische, ökonomische und soziale Interessen die Gestaltung von Softwareprozessen undurchsichtig machen können. Ein häufiges Beispiel, das auch in den FAT/ML Principles erwähnt wird und das Problem ausreichend anerkennt, ist das Abwägen zwischen vollkommener Transparenz eines Systems und dem Schutz des Daten- und Industriegeheimnisses.

Ein Gütekriterienkatalog für gesellschaftlich sensible Algorithmeneinsätze darf nicht lediglich eine Sammlung von technischen Qualitätsmerkmalen wie zum Beispiel der Überprüfbarkeit eines Systems oder der Reduzierung seiner Fehlerquoten sein. Auch die moralische Qualität eines Algorithmus muss berücksichtigt werden. Dazu gehören komparative Konzepte wie Fairness, welches in detaillierter Form in den FAT/ML Principles beschrieben wird, aber auch grundsätzlichere ethische Fragen. Beispielhaft in dieser Hinsicht sind die Asilomar AI Principles, in die unter anderem die Einhaltung der Menschenrechte aufgenommen wurde. Aber auch die dort beschriebene Vereinbarkeit algorithmischer Systeme mit weiteren menschlichen Werten, wie Würde, Rechten, Freiheiten und kultureller Diversität, ist unverzichtbar. Die Behandlung von Metafragen der Einstellung zu künstlicher Intelligenz bzw. deren grundsätzliche Ziele sind ebenfalls in den Asilomar AI Principles vorbildlich eingearbeitet.

Zuletzt ist anzumerken, dass ein detailliertes Verständnis einzelner Gütekriterien zuträglich ist. Da in der Diskussion um die Ethik der Algorithmen häufig wiederkehrende Buzzwords wie Accountability, Transparency oder Auditability gebraucht werden, ist es für die Adressaten von großer Bedeutung, dass diese ausreichend definiert werden. Sicherlich ist die Genauigkeit, mit der beispielsweise in den FAT/ML Principles der Aspekt der Accountability behandelt wird, nicht für alle Merkmale richtungweisend. Dennoch ist auf der anderen Seite die Unterscheidung zwischen Judicial Transparency und Failure Transparency (Asilomar AI Principles) ein hilfreiches Beispiel für ein differenziertes Verständnis einzelner Kriterien. Für praktische Arbeit an einem deutschsprachigen Gütekriterienkatalog bedeutet dies, dass ein detailliertes Verständnis einzelner Kriterien zwar wünschenswert, im Sinne der Anwenderfreundlichkeit jedoch auf kurze Definitionen zu reduzieren ist.

In Bezug auf die formale Gestaltung eines Gütekriterienkataloges lässt sich festhalten, dass Listenformate einen grundsätzlichen Vorteil der Lesbarkeit bieten – dem wird in allen drei analysierten Katalogen Rechnung getragen. Darüber hinaus stellt sich jedoch die Frage, wie die Anwendbarkeit des Gütekriterienkataloges am besten gewährleistet werden kann. Auf Grundlage der Analyse lassen sich zwei erfolgreiche Strategien ableiten: ein gesetzesähnlicher Forderungskatalog oder aber eine Sammlung konkreter Schritte. Ersteres Charakteristikum ist besonders bei den Asilomar AI Principles verwirklicht. Die Prinzipien sind sowohl generell als auch mit einer idealistischen Ausrichtung auf hohe Ziele formuliert, die eine Auslegung in Zweifelsfällen erlauben. Wie bereits erläutert, bietet dies den Vorteil, dass die Diversität des gesamten algorithmischen Gestaltungsprozesses in den Gütekriterien erfasst werden kann und das Dokument nicht lediglich spezielle Anwendungsfelder beschreibt.

Eine zweite Möglichkeit, besonders repräsentiert durch das Social Impact Statement in den FAT/ML Principles, ist die Gewährleistung der Anwendbarkeit durch die Angabe konkreter Schritte. Auch in der Sektion Research Funding der Asilomar AI Principles wird dies ansatzweise realisiert. Das Formulieren konkreter erster Schritte zur Einhaltung bzw. zum Nachweis der Einhaltung der Gütekriterien erleichtert den Verantwortlichen die Umsetzung, schafft Anreize und den Eindruck von Bewältigbarkeit des Vorhabens. Ob bei der Erstellung eines konkreten Gütekriterienkataloges die Entscheidung zugunsten des Formats eines gesetzähnlichen Textes oder einer Sammlung von Handlungsvorschriften entschieden werden sollte, kann nicht abschließend geklärt werden. Es ist auch in Erwägung zu ziehen, beide Ansätze zu kombinieren.

6.2 Auszugleichende Schwächen

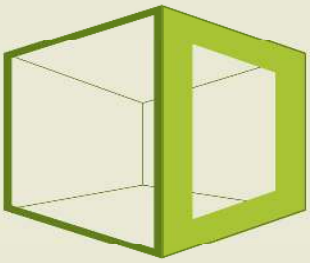
Allen drei analysierten Gütekriterienkatalogen ist gemeinsam, dass sie keine Anhaltspunkte zur praktischen Implementierung des Dokumentes bzw. zur Stärkung von dessen Verbindlichkeit liefern. Praktische Empfehlungen beziehen sich bestenfalls auf die Umsetzung einzelner Kriterien, nicht jedoch darauf, wie die Sammlung der Qualitätskriterien als Ganzes zum neuen Standard werden kann. Streng genommen sind solche Fragen der Implementierung nicht Teil der Gütekriterien und doch fehlt ohne den Bezug darauf ein wichtiger Impuls für die ethische Gestaltung algorithmischer Prozesse. Auch die Rolle der Politik wird in diesem Zusammenhang in allen drei Dokumenten weitestgehend vernachlässigt.

Worauf die Verbindlichkeit erfolgreicher Professionsethiken aus anderen Berufsfeldern beruht, ist wissenschaftlich kaum erforscht, was die Abwesenheit solcher Bezüge in den bestehenden Dokumenten erklären kann. Innerhalb

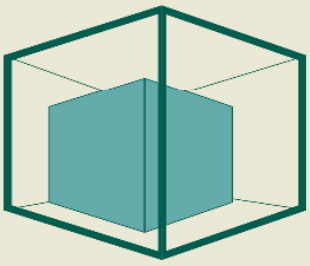
des Projektes Ethik der Algorithmen der Bertelsmann Stiftung wird dieser Frage aktuell durch ein Autorenteam um Prof. Dr. Alexander Filipović nachgegangen und in den kommenden Wochen veröffentlicht. Auch diese Ergebnisse sollten in einen deutschsprachigen Gütekriterienkatalog einfließen. In jedem Fall sollte eine solche Empfehlung für konkrete nächste Schritte zur Implementierung des Dokumentes enthalten, eventuell zugeschnitten auf bestimmte Gruppen von Verantwortlichen wie zum Beispiel politische Entscheidungsträger, öffentliche Institutionen und Entwickler.

Eng verknüpft mit dem Aspekt der Verbindlichkeit ist die Frage nach Verboten. In keinem der oben genannten Gütekriterienkataloge wird das Mittel des Verbotes bei Nichteinhaltung bestimmter Anforderungen an ein algorithmisches System in Betracht gezogen. Diese Vernachlässigung stellt ein häufiges Phänomen in der Debatte um algorithmische Entscheidungsfindung und künstliche Intelligenz dar (vgl. Krüger und Lischka 2018). Ziel eines Gütekriterienkataloges sollte es sein sicherzustellen, dass Systeme, die den darin beschriebenen Anforderungen nicht genügen, nicht zum Einsatz kommen. Dieser Zusammenhang muss aus dem Dokument ersichtlich werden, da die Abwesenheit eines Verweises auf mögliche Verbote die Frage der Sanktionierung und somit auch die der Verbindlichkeit der Gütekriterien offenlässt. Insbesondere wo die ethische Angemessenheit eines Softwaresystems nicht gewährleistet werden kann (z. B. bei drohender Unterwanderung der moralischen Werte einer Gesellschaft oder der Menschenrechte), sollte ein kategorisches Verbot gegen den Produktiveinsatz ausgesprochen werden können.

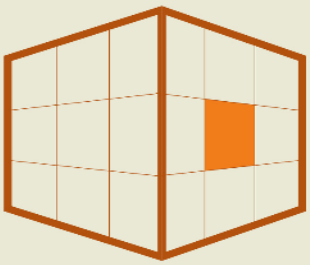
Gütekriterienkataloge für Algorithmen im Vergleich



Formelle Merkmale	ACM US Public Policy Council	FAT/ML Organisation	Future of Life Institute
Listenform	✓	✓	✓
Großer Adressatenkreis (nicht nur Programmierer)	✓	×	✓
Viele Akteure an der Genese beteiligt	✓	×	✓



Inhaltliche Merkmale			
Grundhaltung gegenüber Algorithmen deutlich	✓	✓	✓
Schritte vor und nach der Programmierung berücksichtigt	✓	✓	✓
Daten als Problemquelle berücksichtigt	✓	✓	×
Verantwortung klar beim Menschen verortet	✓	×	✓
Öffentlichkeit als wichtiger Beteiligter berücksichtigt	✓	✓	×
Interessenkonflike berücksichtigt	✓	✓	×
Detailliertes Verständnis einzelner Kriterien vorhanden	×	✓	✓
Forschung als wichtiger Verantwortlicher berücksichtigt	×	×	✓
Ethische Fragen berücksichtigt	×	×	✓
Verbote gegen bestimmte Einsätze ausgesprochen	×	×	×



Implementierung			
Anwendbarkeit ermöglicht durch a) Praktische Anweisungen	×	✓	✓
Anwendbarkeit ermöglicht durch b) Auslegung hoher Ideale	×	×	✓
Nächste Schritte zur Etablierung des Katalogs	×	×	×
Politik als wichtiger Akteur in Implementierung berücksichtigt	×	×	×
Längerfristige Fragen zur Schaffung von Verbindlichkeit berücksichtigt	×	×	×

| BertelsmannStiftung

7 Literatur

Association for Computing Machinery (ACM) (2018). „ACM U.S. Public Policy Council“. https://www.acm.org/public-policy/usacm (Download 14.3.2018).

Association for Computing Machinery (ACM) (2017). „Statement on Algorithmic Transparency and Accountability and Principles for Algorithmic Transparency and Accountability“. https://www.acm.org/binaries/content/assets/public-policy/2017_joint_statement_algorithms.pdf (Download 14.3.2018).

Dreyer, Stephan, und Wolfgang Schulz (2018). "Was bringt die Datenschutz-Grundverordnung für automatisierte Entscheidungssysteme?" Hrsg. Bertelsmann Stiftung. Gütersloh.

Fairness, Accountability, and Transparency in Machine Learning (2018). „Principles for Accountable Algorithms and a Social Impact Statement for Algorithms“. https://www.fatml.org/resources/principles-for-accountable-algorithms (Download 14.3.2018).

Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) (2016). „Dagstuhl Seminar 16291. Data, Responsibly“. https://www.dagstuhl.de/en/program/calendar/semhp/?semnr=16291 (Download 14.3.2018).

Filipovic, Alexander, Claudia Paganini und Christopher Koska (2018). „Verbindlichkeit erfolgreicher Professionsethiken und Übertragbarkeit auf gesellschaftlich relevante algorithmische Prozesse“. Unveröffentlicht. Hrsg. Bertelsmann Stiftung. Gütersloh.

Fischer, Sarah, und Thomas Petersen (2018). "Was Deutschland über Algorithmen weiß und denkt Ergebnisse einer repräsentativen Bevölkerungsumfrage". Hrsg. Bertelsmann Stiftung. Gütersloh.

Future of Life Institute (2018a). „Existential Risk“. https://futureoflife.org/background/existential-risk/. (Download 14.3.2018).

Future of Life Institute (2018b). „The Future of Life Institute (FLI) “. https://futureoflife.org/team/ (Download 14.3.2018).

Future of Life Institute (2017a). „Asilomar AI Principles“. https://futureoflife.org/ai-principles/ (Download 14.03.2018).

Future of Life Institute (2017b). „Beneficial AI 2017“. https://futureoflife.org/bai-2017/ Download (14.3.2018).

Future of Life Institute (2017c). „A Principled AI Discussion in Asilomar“. https://futureoflife.org/2017/01/17/principled-ai-discussion-asilomar/ (Download 14.3.2018).

Krüger, Julia, und Konrad Lischka (2018). „Damit Maschinen den Menschen dienen. Lösungsansätze, um algorithmische Entscheidungen in den Dienst der Gesellschaft zu stellen“. Unveröffentlicht. Hrsg. Bertelsmann Stiftung. Gütersloh.

Lischka, Konrad, und Christian Stöcker (2017). "Digitale Öffentlichkeit". Hrsg. Bertelsmann Stiftung. Gütersloh.

Lischka, Konrad, und Anita Klingel (2017). "Wenn Maschinen Menschen bewerten". Hrsg. Bertelsmann Stiftung. Gütersloh.

Rohde, Noëlle (2017). „In Australien prüft eine Software die Sozialbezüge – und erfindet Schulden für 20.000 Menschen“. *Algorithmenethik.de* 25.10. https://algorithmenethik.de/2017/10/25/in-australien-prueft-eine-software-die-sozialbezuege-und-erfindet-schulden-fuer-20-000-menschen/ (Download 15.3.2018).

Vieth, Kilian, und Ben Wagner (2017). "Teilhabe, ausgerechnet". Hrsg. Bertelsmann Stiftung. Gütersloh.

Zweig, Katharina (2018): "Wo Maschinen irren können. Fehlerquellen und Verantwortlichkeiten in Prozessen algorithmischer Entscheidungsfindung". Hrsg. Bertelsmann Stiftung. Gütersloh.

8 Über die Autorin

Noëlle Rohde ist Projektmanagerin im Team „Ethik der Algorithmen“ der Bertelsmann Stiftung, und dort verantwortlich für das Professionsethik-Modul. Sie studierte Medical Anthropology (M.Sc.) an der University of Oxford und verfasste dort ihre Abschlussarbeit zur Thematik der quantifikatorischen Diskriminierung in Global Health Metrics und Self-Tracking Technologien. Zuvor studierte sie Philosophie, Psychologie und Linguistik in Oxford und Paderborn.

9 Impulse Algorithmenethik

Alle Veröffentlichungen sind abrufbar unter: https://algorithmenethik.de/impulse/

Impuls Algorithmenethik #1: Lischka, Konrad, und Anita Klingel (2017). *Wenn Maschinen Menschen bewerten.* Hrsg. Bertelsmann Stiftung. Gütersloh. (Auch online unter https://doi.org/10.11586/2017025, Download 30.05.2018.)

Impuls Algorithmenethik #2: Vieth, Kilian, und Ben Wagner (2017). *Teilhabe, ausgerechnet.* Hrsg. Bertelsmann Stiftung. Gütersloh. (Auch online unter https://doi.org/10.11586/2017027, Download 30.05.2018)

Impuls Algorithmenethik #3: Lischka, Konrad, und Christian Stöcker (2017). *Digitale Öffentlichkeit.* Hrsg. Bertelsmann Stiftung. Gütersloh. (Auch online unter https://doi.org/10.11586/2017028, Download 30.05.2018)

Impuls Algorithmenethik #4: Zweig, Katharina Anna (2018). *Wo Maschinen irren können. Fehlerquellen und Verantwortlichkeiten in Prozessen algorithmischer Entscheidungsfindung.* Hrsg. Bertelsmann Stiftung. Gütersloh. (Auch online unter https://doi.org/10.11586/2018006, Download 30.05.2018)

Impuls Algorithmenethik #5: Dreyer, Stephan, und Wolfgang Schulz (2018). *Was bringt die Datenschutz-Grundverordnung für automatisierte Entscheidungssysteme?* Hrsg. Bertelsmann Stiftung. Gütersloh. (Auch online unter https://doi.org/10.11586/2018011, Download 30.05.2018)

Impuls Algorithmenethik #6: Lischka, Konrad, und Julia Krüger (2018). *Damit Maschinen den Menschen dienen. Lösungsansätze, um algorithmische Prozesse in den Dienst der Gesellschaft zu stellen.* Hrsg. Bertelsmann Stiftung. Gütersloh. (Auch online unter https://doi.org/10.11586/2017028, Download 30.05.2018)

Impuls Algorithmenethik #7: Fischer, Sarah, und Thomas Petersen (2018). *Was Deutschland über Algorithmen weiß und denkt Ergebnisse einer repräsentativen Bevölkerungsumfrage.* Hrsg. Bertelsmann Stiftung. Gütersloh. (Auch online unter https://doi.org/10.11586/2018022, Download 30.05.2018)

Adresse | Kontakt

Bertelsmann Stiftung
Carl-Bertelsmann-Straße 256
33311 Gütersloh
Telefon +49 5241 81-0

Ralph Müller-Eiselt
Senior Expert Taskforce Digitalisierung
Telefon +49 5241 81-81456
ralph.müller-eiselt@bertelsmann-stiftung.de

Noëlle Rohde
Project Manager
Telefon: +49 5241 81-81141
noelle.rohde@bertelsmann-stiftung.de

www.bertelsmann-stiftung.de

BertelsmannStiftung