Sachdokumentation:

Signatur: DS 3876

Permalink: www.sachdokumentation.ch/bestand/ds/3876

Zitierweise für graue Literatur

Elektronische Broschüren und Flugschriften (DS) aus den Dossiers der Sachdokumentation des Sozialarchivs werden gemäss den üblichen Zitierrichtlinien für wissenschaftliche Literatur wenn möglich einzeln zitiert. Es ist jedoch sinnvoll, die verwendeten thematischen Dossiers ebenfalls zu zitieren. Anzugeben sind demnach die Signatur des einzelnen Dokuments sowie das zugehörige Dossier.

DIGITALE | GESELLSCHAFT

Position der Digitalen Gesellschaft zur Regulierung von automatisierten Entscheidungssystemen

21. Februar 2022

Version 1.0

Erarbeitet von der Fachgruppe ADMS:

David Sommer, Andreas Geppert, Christian Sigg, Daniel Donatsch, Erik Schönenberger, Nicole Pauli und Tanja Klankert

Inhalt

1. Einleitung

Systeme für automatisierte Entscheidungen (Automated Decision Making Systems, ADM-System, ADMS, siehe Glossar) sind in unserem Alltag angekommen. Die Algorithmen (siehe Glossar) sozialer Medien und der Newsportale entscheiden etwa, welche Nachrichten wir sehen. Risikobewertungen haben einen Einfluss darauf, ob und zu welchen Konditionen wir Kredite und Versicherungsleistungen bekommen. Die Vorhersagen eines Predictive-Policing-Systems können angeben, wo Polizeistreifen patrouillieren sollen und wo nicht. Diese ADM-Systeme haben somit einen signifikanten Einfluss auf den Alltag und das Leben von Menschen. Sie können unter Umständen Entwicklungschancen von Individuen beeinträchtigen und sogar ihre Grundrechte verletzen. Darüber hinaus können ADM-Systeme einen gesellschaftlichen Einfluss haben, beispielsweise aufgrund eines algorithmischen «Bias» (siehe Glossar), welcher unter Umständen bestimmte Personen oder Gruppen benachteiligt[1].

Da ADM-Systeme neben dem positiven Nutzen auch negative Auswirkungen auf Individuen und die Gesellschaft haben können, ergeben sich eine Reihe von Fragen. Sind die von ADM-Systemen gefällten Entscheidungen fair und chancengerecht, oder diskriminieren sie bestimmte Individuen oder Personengruppen? Wie können automatisiert getroffene Entscheidungen nachvollzogen werden? Wie kann überprüft werden, ob die getroffene Entscheidung gerechtfertigt ist? Gibt es einzelne Entscheidungen oder ganze Bereiche (z.B. Gerichtsurteile), die wir nicht Algorithmen überlassen sollten, sondern die nur von Menschen gefällt werden sollten? Aus diesen Überlegungen schliessen wir, dass die möglichen Einflüsse von ADM-Systemen auf Individuen wie auf die Gesellschaft einer kritischen Reflexion unterzogen werden müssen und dass sich unsere Gesellschaft Regeln geben muss, in welchen Bereichen und auf welche Art ADM-Systeme eingesetzt werden sollten. Dies ist der gesellschaftliche Kontext, in dem sich der vorliegende Vorschlag für einen Rechtsrahmen bewegt. Eine Regulierung soll auch sicherstellen, dass Nutzen und Risiken in einem guten Verhältnis zueinander stehen. Diesen Rechtsrahmen stellen wir in diesem Beitrag vor.

Eine wesentliche Grundlage für den Umgang mit ADM-Systemen ist die Schaffung von Transparenz. Betroffene Individuen, berechtigte NGOs sowie eine Aufsichtsbehörde sollen ein Recht auf Einsicht für Anwendungen enthalten, die ADM-Systeme einsetzen. Dem Rechtsrahmen liegt eine Einschätzung der Risiken zugrunde, welche von ADM-Systemen ausgehen. Durch die Schaffung von Transparenz wird die Einschätzung des Risikos einer Anwendung ermöglicht. Abhängig vom jeweiligen Risiko gibt es weniger oder mehr Einschränkungen und Regeln, die für das jeweilige ADM-System zu berücksichtigen sind.

1 Z.B. bei der automatisierten Einschätzung der Arbeitsmarkt-Reintegrationschancen von arbeitslosen Personen.

Unser Vorschlag ist technologieneutral[2] und folgt einem «human-centered» Ansatz: ADM-Systeme sollen dem Menschen nutzen, d.h. es soll dem Menschen durch den Einsatz von ADM-Systemen besser gehen[3]. Der Rechtsrahmen trägt auch dem Umstand Rechnung, dass das Risiko eines Systems sich mit der Zeit verändern kann.

2 Im Fokus unseres Regulierungsvorschlags stehen die Auswirkungen und Risiken der ADM-Systeme und nicht Verbote konkreter Technologien.
3 Zugrunde liegt das Prinzip der Benefizienz.

2. Geltungsbereich

In den Anwendungsbereich unseres Vorschlags fallen ADM-Systeme, die Entscheidungen mit Hilfe von technischen Systemen vollständig automatisiert treffen oder zumindest unterstützen. Wir übernehmen die folgende Definition aus einer Empfehlung des AI Now Instituts (Richardson et al. 2019, S. 20) an die Stadt New York:

> An «automated decision system» is any software, system, or process that aims to automate, aid, or replace human decision-making. Automated decision systems can include both tools that analyze datasets to generate scores, predictions, classifications, or some recommended action(s) that are used by agencies to make decisions that impact human welfare[4], and the set of processes involved in implementing those tools.

ADM-Systeme benutzen für die Entscheidungsfindung Algorithmen und/oder Techniken der Künstlichen Intelligenz und sind oft (aber nicht immer) datengetrieben. Dies bedeutet jedoch nicht, dass jeder Algorithmus bzw. jedes KI- oder Big-Data-System unter den Geltungsbereich dieses Gesetzes fallen sollte. Weiterhin ist die Aussage, dass fast jedes Computerprogramm ständig Entscheidungen trifft, zwar prinzipiell richtig, aber aus der Sicht der Regulierung nicht zielführend. Die unter die Regulierung fallenden Entscheidungen müssen als einzelne, diskrete Entscheidungen wahrnehmbar und von einer gewissen Signifikanz sein. Weiterhin müssen sie potentielle Auswirkungen auf Individuen und/oder die Gesellschaft haben, um in den Geltungsbereich dieses Gesetzes zu fallen.

Fällt ein technisches System nicht in den Anwendungsbereich des Rechtsrahmens, so ist keine Risiko-Kategorisierung gemäss Kapitel 6 nötig, und der damit verbundene Aufwand muss nicht geleistet werden.

4 Impact on public welfare includes but is not limited to decisions that affect sensitive aspects of life such as educational opportunities, health outcomes, work performance, job opportunities, mobility, interests, behavior, and personal autonomy.

3. Zusammenfassung Rechtsrahmen

Der Rechtsrahmen folgt einer Mischform zwischen schadens- und risikobasiertem Ansatz. Beim ersten Ansatz werden Sanktionen erst nachträglich im Schadensfall verhängt, während beim zweiten Ansatz risikoreiche Anwendungen von vornherein bestimmten Auflagen unterliegen. Wer ein ADM-System einsetzt, muss dessen Risiko für die Gesundheit, Sicherheit und Grundrechte von Individuen und der Gesellschaft einschätzen und kategorisieren. Der Rechtsrahmen stellt dazu drei Kategorien zur Verfügung: «tiefes Risiko», «hohes Risiko» und «inakzeptables Risiko». Grundsätzlich richten sich die Kategorien nach dem vom System ausgehenden Risiko für Einzelpersonen sowie für die Gesellschaft als Ganzes. So geht von «Systemen mit tiefem Risiko» ein geringes Risiko für Individuen aus und keines für die Gesellschaft, während «Systeme mit inakzeptablem Risiko» ein unvertretbar hohes Risiko für Individuen oder die Gesellschaft darstellen. Dazwischen finden sich die «Systeme mit hohem Risiko». Für diese Systeme gilt eine weitgehende Transparenz- und Sorgfaltspflicht, welche für die Öffentlichkeit die Einschätzung des Risikos und damit deren Nutzen ermöglichen sollen. Denn im Gegenzug zu Systemen mit einem inakzeptablen Risiko werden jene mit einem hohen Risiko nicht verboten.

In erster Linie betrachten wir die Auswirkung von ADM-Systemen auf Individuen. Jedoch kann bei breitem Einsatz von einfach skalierenden Systemen auch ein Risiko für die Gesellschaft[5] entstehen, das nur mit Blick auf Individuen nicht ausreichend messbar oder sanktionierbar ist. ADM-Systeme, die individualisierte politische Werbung in sozialen Netzwerken massenhaft verteilen, sind ein Beispiel für ein derartiges gesellschaftliches Risiko. Im Einzelfall kann das Risiko mit Verweis auf die individuelle Souveränität vernachlässigbar sein. Im Mittel über viele Menschen können solche ADM-Systeme jedoch merkliche Auswirkungen, etwa auf Wahlresultate, haben und somit der Demokratie Schaden zufügen. Die bisherige auf Individuen fokussierte Schweizer Rechtsprechung, wie beispielsweise jene zum Datenschutzgesetz, greift in solchen Fällen zu kurz. Unser Vorschlag geht dieses Defizit an, in dem er das Risiko für die Gesellschaft anerkennt und mit kollektiven Rechtsbehelfen, wie Sammelklagen und einem Verbandsklagerecht, Abhilfe schafft.

Der Rechtsrahmen unterscheidet zwischen ADM-Systemen, welche in der Privatwirtschaft eingesetzt werden, und solchen, die in Erfüllung öffentlicher Aufgaben verwendet werden. Für beide fordern wir ein Beschwerde-, resp. Klagerecht für betroffene Individuen, die staatliche ADM-Aufsichtsbehörde und berechtigte NGOs, um die korrekte Risiko-Klassifizierung gemäss Kapitel 6 und die Durchsetzung der damit verbundenen Pflichten zu garantieren.

Die neu zu schaffende ADM-Aufsicht (mehr dazu im Kapitel 5.1) soll Reklamationen sammeln, auf Verdacht den Einsatz von ADM-Systemen in Unternehmen und staatlichen Stellen überprüfen sowie erstinstanzlich umsatzabhängige Verwaltungssanktionen verhängen können. Als diverses Fachgremium, zusammengesetzt aus Personen mit sozialwissenschaftlicher, techni-

5 https://booksummaryclub.com/weapons-of-math-destruction-book-summary/

scher und juristischer Expertise, soll sie finanziell und personell unabhängig und frei von Weisungen agieren.

Die genaue Funktionsweise eines ADM-Systems unterliegt meist dem Geschäftsgeheimnis. Entsprechend schwierig ist es für Aussenstehende, Beweise für das Risiko eines Systems zu beschaffen. Daher soll eine Beweislastumkehr gelten. Das heisst ein Unternehmen muss die korrekte Klassifizierung nachweisen, sofern ein Gericht auf eine Klage eintritt. Für Systeme in Erfüllung eines öffentlichen Auftrags fordern wir weitgehende Transparenz und Veröffentlichung der Systeme und Daten (siehe Glossar), ganz im Sinne der Forderung «Public Money? Public Code!»[6]. Ausnahmen, beispielsweise für Personendaten, werden weiter unten aufgelistet.

Die staatliche ADM-Aufsicht unterstützt die Verantwortlichen bei der Risiko-Einschätzung von ADM-Systemen durch Checklisten und Good-Practices-Anleitungen, um für Sensibilisierung und adäquate Handhabung zu sorgen. Wer sein eingesetztes System falsch einschätzt und dadurch seinen Pflichten nicht nachkommt oder ein verbotenes, mit inakzeptablem Risiko behaftetes ADM-System betreibt, dem sollen empfindliche und umsatzabhängige Strafen drohen. Dabei soll es sich um Verwaltungssanktionen handeln, die explizit nicht auf die Bestrafung einzelner Mitarbeiter durch das Strafrecht abzielen, da es sich in der Regel nicht um ein individuelles, sondern um ein Organisationsverschulden handelt. Diese Sanktionen sollen jedoch das letzte Mittel bleiben.

Um Innovation nicht zu verhindern, setzen wir auf Selbstdeklaration, statt die Unternehmen und die öffentliche Verwaltung mit bürokratischen Prüfprozessen zu belasten. Dies erlaubt den betroffenen Branchen und Akteuren, die konkrete Umsetzung zur Einhaltung der Regeln innerhalb der durch den Rechtsrahmen definierten Parameter selbst zu gestalten.

Eine zunehmende Anzahl an international bedeutenden Institutionen, wie die Europäische Union oder der amerikanische Berufsverband der Informatiker:innen (Association for Computing Machinery, ACM), beschäftigt sich mittlerweile mit der Notwendigkeit einer Regulierung von ADM-Systemen (siehe Anhang Kapitel B). Wir sind überzeugt, dass der vorgeschlagene Rechtsrahmen dazu beiträgt, die existierenden Regulierungslücken zu schliessen. Im Folgenden werden die Kernaspekte des Rechtsrahmens – insbesondere die Risikokategorien und die Transparenz- und Sorgfaltspflichten – im Detail erläutert.

6 Von der Öffentlichkeit bezahlte Software und deren Source Code soll offen und für alle zugänglich sein. https://publiccode.eu/

4. Die gesellschaftliche Relevanz

Dieses Kapitel erläutert, weshalb sich die Digitale Gesellschaft für ein ADM-Gesetz einsetzt.

Automatisierte Entscheidungssysteme sind weder objektiv noch neutral. Diese Einsicht ist für die Relevanz des vorliegenden Rechtsrahmens zentral. Zum einen tragen Systeme stets die Werte ihrer Entwickler:innen sowie jene der Gesellschaft mit. Zum anderen sind sie durch ihre Funktion gebunden: Von Entwickler:innen bewusst oder unbewusst getroffene Design-Entscheide haben einen Effekt auf die Wirkungsweise des Systems und können negativer Natur sein. Die Funktion eines Systems bestimmt einen engen Handlungsrahmen, der oftmals nicht hinterfragt wird.

Entsprechend können automatisierte Systeme auch als sozio-ökonomischer Spiegel einer bestimmten Gesellschaft betrachtet werden. Problematisch wird dies auch deshalb, weil sich kulturelle und gesellschaftliche Werte rund um den Globus unterscheiden können, wohingegen Technologien wie ADM-Systeme grenzüberschreitend oder global Einsatz finden. Das Bewusstsein für diese Problematik ist jedoch noch zu wenig in der Gesellschaft verbreitet.

Die zunehmende Unterstützung durch ADM-Systeme hat viele Vorteile. Sie werden aufgrund ihrer Nützlichkeit aber oftmals als objektive Helfer wahrgenommen, deren Entscheide zwingend richtig sein müssen. Die Ursache dafür ist «**Technologiegläubigkeit**»: Der Mensch vertraut der Maschine, dass deren Rechenresultate objektiv und korrekt sind.

Dieses Gefühl von Sicherheit und Objektivität ist jedoch trügerisch, denn für viele Entscheidungsprobleme existiert keine optimale Lösung. Entscheidungen von gesellschaftlicher Tragweite sind Gegenstand von Aushandlungsprozessen und können international variieren. Dabei kann das Delegieren alltäglicher Aufgaben an ein ADM-System dazu führen, dass die Interaktion mit Maschinen sowie ihren Resultaten Teil der sozialen Realität werden. Die Soziologin Michele Willson schreibt in ihrem Essay: «Einem Algorithmus wird eine Aufgabe oder ein Prozess übertragen, und die Art und Weise, wie er eingesetzt wird und mit ihm umgegangen wird, wirkt sich wiederum auf die Dinge, Menschen und Prozesse aus, mit denen er interagiert - mit unterschiedlichen Folgen» (Willson 2017: 139). Dieser Rückkopplungseffekt führt dazu, dass automatisierte Entscheidungssysteme sowie ihre (teils fehlerhaften oder ungenauen) Datengrundlagen sich stets verändern und Teil des sozialen Gefüges werden. Die Abgabe von Aufgaben an automatisierte Systeme kann auch dazu führen, dass individuelle und kollektive Verantwortlichkeit abnehmen.

Effekte automatisierter Entscheidungssysteme können sowohl beabsichtigt wie unbeabsichtigt sein. Dies ist besonders problematisch im Falle von Diskriminierung. Zum Beispiel reproduzieren auf Datensätzen trainierte ADM-Systeme die darin implizit festgeschriebenen Diskriminierungspraktiken. So würden auf historischen Daten trainierte Recruitment-Systeme Frauen vermutlich nach wie vor häufiger benachteiligen, obwohl dies mittlerweile nicht mehr toleriert wird. Menschen treffen zwar nicht per se die besseren Entscheidungen und sind nicht frei von Vorurteilen. Sie können diese aber, auch mit Hilfe von Technologien, reflektieren und sich

darüber austauschen. Diese **Reflexionsfähigkeit fehlt den Systemen**, weshalb ihnen keine Entscheidungen delegiert werden sollten, welche nachhaltige Auswirkungen auf die Gesellschaft haben. Diskriminierende Effekte von ADM-Systemen können sich vor allem in Kombination mit der Technologiegläubigkeit durch den vermehrten, auch grenzüberschreitenden Einsatz verstärken.

Automatisierte Entscheidungssysteme kuratieren zunehmend den Informationsüberfluss, zum Beispiel in Form von sogenannten «Vorschlagssystemen» oder als «Fakten-» und «Urheberrechtsprüfer». So werden Systembetreiber ermächtigt, politische Botschaften und Positionen durch gezieltes Agenda-Setting selektiv zu verstärken. Dabei müssen diese Systeme (z.B. als Faktenprüfer) nicht zwingend auf Personendaten operieren. Viele dieser Effekte von automatisierten Entscheidungssystemen ist gemein, dass sie oft im Verborgen geschehen und ihre Auswirkungen erst spät oder durch weitere, indirekte Effekte bemerkt werden. Einsatz und Funktionsweise der Systeme ist oftmals nicht bekannt, da ihre Entwicklerinnen und Betreiber kein Interesse an einer Offenlegung haben.

Die zunehmende Verarbeitungsgeschwindigkeit und Übermittlungsmöglichkeiten von Informationen erlauben eine stärkere Vernetzung als bisher möglich war. Das Zusammenspiel unterschiedlicher ADM-Systeme lässt dabei zusätzliche Risiken entstehen, die nur schwer abzuschätzen sind. Die Entstehung von sich durch Rückkopplung verstärkenden Effekten (Feedback-Loops, siehe Glossar) ist absehbar und damit ein gesellschaftliches Risiko (mehr dazu im Anhang Kapitel A). Dieser **Steigerung der Komplexität** stehen aber immer noch Menschen gegenüber, welche zunehmend Schwierigkeiten haben, diese, selbst mit voller Transparenz der individuellen Systeme, zu durchdringen.

Um diese Effekte ansatzweise abschätzen zu können, sollte daher eine weitreichende **Transparenz- und Sorgfaltspflicht** gelten. Es sollte zumindest bei wichtigen automatisierten Entscheidungssystemen ein Raum für einen nachhaltigen, öffentlichen Diskurs über die Normen und Werte geschaffen werden, die den angelegten, ausgewerteten und interpretierten Metriken, Mess- oder Kennzahlen zugrunde liegen. **Der Mensch soll die Geltungshoheit über ADM-Systeme besitzen** und nicht umgekehrt.

Weiter sind viele Effekte aus Einzelperspektive nicht klar fassbar und werden erst durch die akkumulierten Betrachtungen vieler Betroffener sichtbar. Leider argumentieren die bestehenden Gesetze jedoch meist aus einer Einzelfallperspektive. Daher benötigen wir Methoden zur **kollektiven Rechtsdurchsetzung**, die bisher erst selten im Schweizer Recht zu finden sind.

Durch den Fokus auf die Auswirkungen und Risiken erlaubt eine **technologieneutrale Formulierung** flexibel auf neue Methoden oder geänderte Einsatzmöglichkeiten bestehender Technologien zu reagieren. Das Ziel sollte sein, dass es dem Mensch durch den Einsatz dieser System besser geht. Diese und weitere Themen werden in den Wissenschaften intensiv diskutiert, in diesem Kapitel des Positionspapiers der Vollständigkeit halber allerdings nur skizziert.

5. Ein Gesetz für automatisierte Entscheidungssysteme (ADM-Gesetz)

Wir werden als Gesellschaft zunehmend mit den spezifischen Auswirkungen von ADM-Systemen konfrontiert. Daher fordern wir **ein eigenes «Gesetz für automatisierte Entscheidungen»** oder zumindest eine substantielle Erweiterung der bestehenden Gesetze. Wir fordern **Transparenz** beim Einsatz von ADM-Systemen, um die bereits bestehen Gesetze anwenden zu können, und **wirksame Strafen** bei Missachtung. Wir fordern **Erklärbarkeit** von ADM-Systemen, welche sich dem Menschen rasch und mit angemessenem kognitiven Aufwand erschliesst. Wir fordern eine rechtsstaatliche Kontrolle kritischer ADM-Systeme mit der Möglichkeit, bei Bedarf eingreifen zu können.

Wir sehen in erster Linie konkrete Auswirkungen auf Individuen und wollen ihnen Möglichkeiten zur Durchsetzung ihrer Rechte geben. Es gibt jedoch Risiken, die eher die Gesellschaft als Ganzes betreffen, so zum Beispiel die politische Einflussnahme durch personalisierte Werbekampagnen oder selbstverstärkende Rückkopplungseffekte, bei denen verkettete Systeme – mit oder ohne menschlichen Zutuns – eigene Wertekreisläufe bilden könnten. **Transparenz ist zentral, aber ohne weitere Massnahmen nicht ausreichend.** Das Recht auf informationelle Selbstbestimmung verlangt zusätzlich zur Kenntnis der Vorgänge auch Möglichkeiten, in gewissem Mass Kontrolle darüber auszuüben.

Wir fordern **klare Schutzziele, nämlich die Einhaltung der Grund- und Menschenrechte, die Wahrung der psychischen und physischen Gesundheit und Sicherheit des Einzelnen, Wahrung der Lebens- und Entwicklungschancen, sowie der Schutz der demokratischen Rechte und Prozesse**. Weiter muss die Öffentlichkeit die Möglichkeit haben, die Einhaltung dieser Schutzziele effektiv zu kontrollieren und zu beanstanden. Der Mensch soll die Geltungshoheit besitzen, er soll also mit seiner Interpretation generell über der Maschine stehen und durch ADM-Systeme seine Ideen und Ziele besser, schneller und weniger fehlerbehaftet erreichen können.

Das ADM-Gesetz soll weder die Innovation hemmen noch die Unternehmen und die später im Detail beschriebene ADM-Aufsicht bürokratisch unverhältnismässig belasten. Wir sprechen uns für einen breiten Rechtsrahmen aus, welche die benötigte Regulierung generell einführt, aber es den einzelnen Wirtschaftssektoren ermöglicht, die effektivsten Methoden zur Umsetzung der Schutzziele selbst zu bestimmen. Unsere später im Detail beschriebene technologieneutrale und risikobasierte Kategorisierung ist mit dem vorgeschlagenen **AI Act der Europäischen Union kompatibel**, erlaubt jedoch im Gegensatz zu der anwendungsbasierten Kategorisierung der EU eine kontextabhängige Einordnung von Anwendungen. Im Anhang Kapitel B findet sich eine Übersicht anderer Regulierungsbemühungen, für die Schweiz und international.

Grundsätzlich können auch bestehende Gesetze mit einigen Änderungen ebenfalls auf ADM-Systeme Anwendung finden. So kann beispielsweise die Dispersion und Verwendung von Per-

sonendaten durch das Datenschutzgesetz geregelt werden[7]. Diskriminierungsverbote sind das Negativ zu der aufkommenden Fairness-Diskussion[8], jedoch mit einem grossen Graubereich dazwischen[9], in dem sich der Hauptteil der realen Anwendungen wiederfinden wird. Das Arbeits- sowie das Datenschutzrecht verbietet einige Überwachungspraktiken, algorithmische und nicht algorithmische, am Arbeitsplatz.[10] Formaljuristisch sollte ein eigenes ADM-Gesetz aus zwei Gründen geschaffen werden: Erstens, weil die nötigen Änderung in anderen Gesetzestexten eine gemeinsame Begriffserklärung sowie Definition der Kategorisierung und der Risiken benötigen und zweitens, weil die im Anschluss ausgeführte ADM-Aufsicht schlecht anderswo definiert werden kann.

5.1 Die ADM-Aufsicht

Die **staatliche ADM-Aufsicht** soll **als Kompetenzzentrum** wirken. Sie berät Unternehmen, Behörden und die Öffentlichkeit und orchestriert allfällige Langzeitanalysen. Sie sammelt Beschwerden von Betroffenen und kontrolliert unabhängig von Weisungen bei hinreichendem Verdacht die Einhaltung des Gesetzes und die Kategorisierung von staatlichen und privatwirtschaftlichen ADM-Systemen.

Die ADM-Aufsicht soll auf allen Ebenen (Bund, Kantone und Gemeinden) massgebend sein. Sie kann parallel zu den Beschwerde- und Klagewegen der Individuen und der berechtigten NGO bei Verstössen erstinstanzlich Sanktionen verhängen. Die Kompetenz, über die die Einhaltung des ADM-Gesetzes und der Verhinderung und Sanktionierung der Risiken für Individuen und der Gesellschaft zu wachen, konzentriert sich auf dieser Behörde, wobei ihr einheitlich und für den gesamten öffentlichen Bereich auf allen Ebenen öffentliche Aufgaben zukommen (siehe Kapitel 8).

---

7 Wir fordern eine Anpassung von Art. 21 Abs. 1 nDSG (streichen von «ausschliesslich»): «Der Verantwortliche informiert die betroffene Person über eine Entscheidung, die auf einer automatisierten Bearbeitung beruht und die für sie mit einer Rechtsfolge verbunden ist oder sie erheblich beeinträchtigt (automatisierte Einzelentscheidung).»

8 Bei Fairness im Bezug auf Entscheidungsalgorithmen geht es um die Evaluierung und Korrektur von algorithmischem Bias (Verzerrungen). Ausgaben von Entscheidungsalgorithmen werden dabei als «fair» angesehen, wenn sie unabhängig von spezifischen Variablen wie Geschlecht, Alter etc. sind. Die genaue (mathematische) Formulierung von Fairness ist jedoch eine noch offene Debatte, und einige Definitionen widersprechen sich sogar.

9 Während die Diskriminierung als Tatbestand eine schwerwiegende Verletzung der Fairness voraussetzt, ist perfekte Fairness meist nur für eine spezifische Metrik möglich und muss andere (ebenso valide Metriken) vernachlässigen. Dazwischen liegt ein grosser Graubereich.

10 Die Überwachung von Arbeitnehmer:innen sind in begrenztem Mass erlaubt, wie z.B. die Erfassung und Einhaltung der Arbeitszeit (Art. 46 ArG), Daten im Zusammenhang mit Eignung für das Arbeitsverhältnis etc.. Die Grenzen der Überwachung liegen im Persönlichkeitsschutz (Art. 328 ff. OR), Datenschutz nach DSG und in gewissen zwingenden Artikeln des Arbeitsgesetzes. Systematische Überwachungen des Verhaltens der Arbeitnehmer:innen sind unzulässig (Art. 26 Abs. 1 ArGV 3), da sie gesundheitliche Auswirkungen für Arbeitnehmende haben können. Ausnahmen können zulässig sein (Art. 26 Abs. 2 ArGV 3), wenn diese aus anderen Gründen erfolgen, wie z.B. zur Optimierung der Leistung oder Qualitätssicherung und nur, wenn Verhältnismässigkeit gewahrt wird und Gefährdung der Persönlichkeit und Gesundheit aufs Geringste beschränkt werden (Einzelfallabwägung)(vgl. Bürgi und Nägeli 2022).

Sie unterstützt die Einschätzung der Risiken von ADM-Systemen durch Checklisten und Good-Practices-Anleitungen, um für Sensibilisierung und adäquate Handhabung rund um die Problematik zu sorgen. Sie soll eine diverse Behörde (zusammengesetzt aus Personen mit unterschiedlicher sozialwissenschaftlicher, technischer und juristischer Expertise) sein, welche unabhängig von Weisungen und mit eigenem Budget, beispielsweise wie die FINMA, agiert. Wie ihre Aufsichtstätigkeit im Sinne der vorliegenden Grundsätze zum bestmöglichen Schutz von Individuen und Gesellschaft konkret umzusetzen wäre, bleibt im Detail noch zu bestimmen.

5.2 Die IT-Sicherheit

Automatisierte Systeme sind, wie auch andere IT-Systeme, nie hundertprozentig sicher und so unter Umständen «hackbar» oder missbrauchbar. Ihre Funktionen können somit durch Insider oder von Dritten potentiell manipuliert werden. Massnahmen zur Unterbindung derartiger Angriffe gehören unserer Ansicht nach jedoch nicht in ein ADM-Gesetz, sondern in ein generelles «**IT-Sicherheits-Gesetz**».

Das ADM-Gesetz soll sich vielmehr mit den spezifischen Auswirkungen von automatisierten Entscheidungssystemen befassen. Die Risikoklassifizierung darf dabei nicht nur die beabsichtigte Verwendung des Systems berücksichtigen, sondern muss auch vorhersehbare, mögliche falsche und missbräuchliche Verwendungen [11] in Betracht ziehen («reasonably foreseeable misuse», EU AI Act Art 9.2(c)). Ein Beispiel ist die Analyse der Kommunikation aller Mitarbeitenden zum Zweck der Verbesserung der Zusammenarbeit. Ein vorhersehbarer Missbrauch dieses Systems ist die Überwachung und/oder Bewertung der Mitarbeitenden.

Ein wichtiger Baustein für die Zuverlässigkeit von Algorithmen ist die noch fehlende Anwendung der **Produkthaftung auf Software** und damit auch auf alle Arten von Computer-Algorithmen. Es sollte nicht möglich sein, dass sich Softwarehersteller durch geschickte Formulierung von AGBs jeglicher Verantwortung entziehen können. Aufgrund ihrer Allgemeinheit sollte die Produkthaftung jedoch Teil eines solchen IT-Sicherheitsgesetzes sein und nicht nur spezifisch für ADM-Systeme definiert werden.

11 Darunter fallen zum Einen das unrechtmässige Verwenden oder «Hacken» dieser Systeme, aber auch ADM-System-spezifische Effekte, wie die Extraktion sensibler Trainingsdaten aus den Modellen (siehe Glossar) selbst, die Unzuverlässigkeit der Vorhersagen bei Datenreihen mit denen man nicht getestet hat (Fragilität), speziell generierte, für den Menschen korrekt aussehende aber in falschen Ausgaben resultierende Datenreihen (Adverserial Examples) etc.

6. Kategorisierung

Unser Vorschlag folgt einer Mischform zwischen einem schadensbasierten und einem risikobasierten Ansatz. Beim schadensbasierten Ansatz werden Sanktionen erst nachträglich im Schadensfall verhängt. Bei einer risikobasierten Regulierung unterliegen Anwendungen von vornherein entsprechenden Auflagen. Als Ergebnis dieser Mischform werden risiko- und auswirkungsreiche Applikationen von vornherein mit Sorgfalts- und Transparenzpflichten belegt, während wir bei weniger risiko- und auswirkungsreichen Applikationen auf Selbstdeklaration setzen. Potentielle Pflichtverletzungen oder Fehlkategorisierungen werden a posteriori durch Strafen im Rahmen von Beschwerden und Klagen geahndet.

Dabei teilen wir ADM-Systeme in drei Kategorien ein: «tiefes Risiko», «hohes Risiko» und «inakzeptables Risiko». Die Einstufung von automatisierten Entscheidungssystemen erfolgt bezüglich ihres Risikos hinsichtlich der Auswirkungen auf Individuen – die Einzelfallperspektive – und die Gesellschaft. Das Risiko für die Gesellschaft wird aufgrund des Schadenspotentials sowie der Eintrittswahrscheinlichkeit eruiert, während im individuellen Einzelfall das Risiko für die Gesundheit, Sicherheit und Grundrechte betrachtet wird. Wir führen diese Risiken im nächsten Abschnitt genauer aus. Danach erläutern wir Beurteilungskriterien, nach denen ADM-Systeme kategorisiert werden sollen. Abschliessend werden die konkreten Kategorien erläutert.

6.1 Risiken für «Gesundheit, Sicherheit und Grundrechte» sowie Risiken für die Gesellschaft

Im Folgenden gehen wir auf die Risiken für die «Gesundheit, Sicherheit und Grundrechte», die sich aus der Anwendung von ADM-Systemen für Individuen ergeben können, und auf Risiken, die die Gesellschaft betreffen können, ein. Für eine detaillierte Diskussion einiger dieser Punkte verweisen wir auf Seite 43ff des «Gutachtens der Datenethikkommission» der Deutschen Bundesregierung.[12] [13]

Für Individuen sowie Personengruppen umfasst das Risiko für «Gesundheit, Sicherheit und Grundrechte» durch ADM-Systeme insbesondere das Risiko der Verletzung des Rechts auf individuelle Selbstbestimmung und -entfaltung: Das Recht also, die individuelle Identität auszubilden, nach aussen zu zeigen und zu ändern, sowie die individuellen Lebensziele und Lebensweise zu bestimmen, und damit die Entfaltung und Darstellung des eigenen Selbst als Ausdruck der Freiheit des Menschen zu gewährleisten.

Diese individuelle Selbstbestimmung ist ein wesentlicher Teil der Menschenwürde. Voraussetzung für die Selbstbestimmung ist der Schutz der Privatheit, der Schutz vor falscher Darstel-

12 vgl. Datenethikkommission der Bundesregierung 2019

13 Eine ausführliche Analyse der Grundrechtsimplikationen von Gesichtserkennungstechnologie ist in (FRA 2019) zu finden.

lung in der Öffentlichkeit, der Schutz vor heimlicher oder falscher Datenerhebung und -verwendung sowie der Schutz vor falscher Einschätzung durch ein ADM-System.

Der Begriff Sicherheit betrifft die Datenerhebung und -verwendung und damit verbunden die Konsequenz aus Fehlfunktionen, potentiellen Angriffen auf ADM-Systeme und Manipulationen in bösartiger Absicht, die negative Auswirkungen auf die körperliche und psychische Sicherheit sowie die Gesundheit der betroffenen Individuen haben können.

Diese benannten Risiken für Individuen oder Personengruppen fassen wir unter den Überbegriffen «**Gesundheit, Sicherheit und Grundrechte**» zusammen. Wir schliessen vom Begriff Sicherheit explizit den Ausbau der polizeilichen und nachrichtendienstlichen Überwachungsmassnahmen aus.

Die zusätzlichen **Risiken für die Gesellschaft** ergänzen obige Feststellungen fliessend, spannen aber weitere Dimensionen auf, welche durch den Einsatz von ADM-Systemen kritisch werden können.

Insbesondere gilt es die Freiheit und Gleichheit demokratischer Willensbildung und Wahlen sowie Abstimmungen vor Manipulation und Radikalisierung durch ADM-Systeme zu schützen. Automatisierung und Machtkonzentration von Medienintermediären mit Gatekeeper-Funktion stellen für die demokratische Willensbildung und die Demokratie eine erhebliche Bedrohung dar. «Chilling-Effects» durch Überwachung (vgl. Penney 2016) und ihre negativen Auswirkungen auf das Wahrnehmen der Grundrechte (vgl. Assion 2014) sind bereits Realität.

Erziehung und Bildung spielen bei der Sicherung einer freiheitlich-demokratischen Grundordnung eine herausragende Rolle, da sie auf vielfältige Weise die für eine Demokratie konstitutive, kritische Beteiligung der Bürgerinnen und Bürger an der Gestaltung der Gesellschaft, das Verständnis und die Einschätzung gesellschaftlich relevanter Zusammenhänge und Entwicklungen und damit auch letztlich das Vertrauen in eine wertebasierte, gestaltbare Zukunft beeinflusst. Auch Gerechtigkeit und das Empfinden von Gerechtigkeit hängt mit der Möglichkeit der Beteiligung an gesellschaftlichen Prozessen zusammen. Dementsprechend weitreichend ist der Einfluss automatisierter Systeme in diesen Bereichen. Und dementsprechend wichtig ist auch die Teilhabe Vieler an den Automatisierungsmechaniken, um damit partizipative Prozesse und die Solidarität zu stärken und einen systemischen Ausschluss grosser Bevölkerungsgruppen zu verhindern.

Weiter sei hier auch das bereits mehrfach genannte Risiko von sich verstärkenden Rückkopplungsschleifen erwähnt: Mit steigender Komplexität und Vernetzung der Systeme steigt das Risiko, dass sich dynamische (Informations-)Kreisläufe ergeben, deren Auswirkungen eine Person zwar ausgesetzt ist, aber nicht direkt an der verantwortlichen Mechanik teilnimmt.

6.2 Beurteilungskriterien

Mittels den nun ausgeführten Risiken schlagen wir eine Kategorisierung vor. Für die Kategorisierung übernehmen und ergänzen wir einige der Richtlinien aus dem AI Act der EU Kommissi-

on (Art 7.2). Bei der Zuweisung eines ADM-Systems in eine Risikokategorie sollten folgende Aspekte berücksichtigt werden:

- Welches sind Zweck und Einsatzbereich des ADM-Systems?
- In welchem Ausmass wird das ADM-System (voraussichtlich) verwendet werden (punktuell oder flächendeckend)?
- In welchem Ausmass wurden durch die Verwendung des ADM-Systems bekanntermassen bereits die Gesundheit geschädigt, die Sicherheit beeinträchtigt oder negative Auswirkungen auf die Grundrechte verursacht? Gibt es aufgrund von Berichten oder dokumentierten Behauptungen, die den zuständigen Behörden übermittelt werden sollten, Anlass zu erheblichen Bedenken hinsichtlich des Eintretens solcher Schäden, solcher Beeinträchtigungen oder solcher nachteiliger Auswirkungen?
- Worin besteht das potenzielle Ausmass solcher Schäden, solcher Beeinträchtigungen oder solcher nachteiliger Auswirkungen, insbesondere hinsichtlich ihrer Intensität und ihrer Eignung, sich auf eine Vielzahl von Personen auszuwirken?
- In welchem Ausmass sind potenziell geschädigte oder beeinträchtigte Personen von dem von einem ADM-System hervorgebrachten Ergebnis abhängig und in welchem Ausmass sind sie auf das ADM-System angewiesen, weil es insbesondere aus praktischen oder rechtlichen Gründen nach vernünftigem Ermessen unmöglich ist, sich dem Einsatz des ADM-Systems zu entziehen?
- In welchem Ausmass sind potenziell geschädigte oder beeinträchtigte Personen gegenüber dem Nutzer eines ADM-Systems schutzbedürftig, insbesondere aufgrund eines Ungleichgewichts in Bezug auf Machtposition, Wissen, wirtschaftlicher oder sozialer Umstände oder des Alters?
- Zu welchem Grad und wie einfach kann das mit einem ADM-System hervorgebrachte Ergebnis rückgängig gemacht werden? Ergebnisse, die sich auf die Gesundheit oder Sicherheit von Personen auswirken, können nicht als leicht rückgängig zu machen gelten.
- Können destruktive oder sich selbst verstärkende Feedback-Loops (siehe Anhang) entstehen und welche Massnahmen werden dagegen getroffen?

Bei der Kategorisierung eines ADM-Systems, also der Zuweisung eines Systems in eine Risikokategorie, müssen alle diese Aspekte in Kombination berücksichtigt werden. Insbesondere sind Vermeidbarkeit und Revidierbarkeit wichtige Aspekte dabei. Kann die Anwendung eines ADM-Systems auf ein Individuum von diesem unter der Voraussetzung von durchschnittlichen Kenntnissen und normalen Umständen sowie ohne die Inkaufnahme von Nachteilen vermieden werden, so fällt dieses System in eine tiefere Kategorie, als wenn ein Individuum von diesem System abhängig ist. Kann der Effekt einer automatisierten Entscheidung (leicht) rückgängig gemacht (oder kompensiert) werden, so fällt dieses System ebenfalls in eine tiefere Kategorie. Voraussetzung dafür ist, dass Individuen nicht nur Kenntnis von der automatisierten Entscheidung an sich haben, sondern auch von der Möglichkeit des Einspruchs und der Rückgängigmachung, und dass diese Rückgängigmachung mit normalerweise anzunehmenden Kenntnissen verlangt werden und ohne Inkaufnahme von Nachteilen zeitnah erfolgen kann.

6.3 Die Kategorien

Fällt ein konkretes technisches System unter den Geltungsbereich dieses Gesetzes (d.h. handelt es sich um ein ADM-System gemäss Kapitel 2), so soll es in eine der folgenden drei Kategorien eingeteilt werden: **«tiefes Risiko»**, **«hohes Risiko»** und **«inakzeptables Risiko»**. Die weiter unten aufgeführten Sorgfalts- und Transparenzpflichten gelten nur für Systeme mit «hohem Risiko».

Bei dieser Einteilung werden die Risiken gemäss Kapitel 6.1 berücksichtigt und die Beurteilungskriterien gemäss Kapitel 6.2 angewandt. Die Einschätzung, ob es sich um ein ADM-System handelt, und welches Risiko damit verbunden ist, wird selbstdeklarativ von der ADM-System-einsetzenden Entität vollzogen. Damit soll der bürokratische Aufwand so klein wie möglich gehalten werden. Bei falscher oder zu tiefer Selbstdeklaration drohen in Abhängigkeit der Schuldhaftigkeit hohe und umsatzabhängige Verwaltungssanktionen. Die Einschätzung wird daher letztlich den Gerichten zufallen.

Diese risikobasierte Kategorisierung ist kompatibel zur anwendungsbasierten Formulierung des AI Act der Europäischen Union (EU AI Act). Im Gegensatz zum EU AI Act verbieten wir jedoch nicht grundlegend Anwendungen, sondern betrachten sie im Licht der jeweiligen Umstände. So kann algorithmische Emotionserkennung zwar bei Einstellungsgesprächen verboten sein, eine Kunstexposition darf sie aber einsetzen, da das Risiko für die Gesellschaft und die Einzelperson im zweiten Fall tief ist. Das Risiko von automatisierten Entscheidungssystemen und damit deren Einschätzung kann sich auch über die Zeit und mit der Entwicklung von Technik und Gesellschaft sowie im Zusammenwirken mit anderen Systemen ändern. Das vorgeschlagene Kategorisierungsschema kann diese Entwicklungen abbilden.

In die unterste Kategorie («Kein/tiefes Risiko») fallen Systeme,

- die ein tiefes Risiko für die Gesellschaft darstellen und
- die für Individuen
    - kein (oder nur ein geringes) Risiko für die Gesundheit, Sicherheit oder Grundrechte darstellen.

Die Systeme in dieser Kategorie sind also dadurch gekennzeichnet, dass sie aller Voraussicht nach keine besonderen negativen Auswirkungen auf Individuen oder die Gesellschaft haben. Sind mittlere Schäden oder Grundrechtseingriffe möglich, kann das System aber leicht und ohne grössere Spezialkenntnisse vermieden werden oder können schädigende Auswirkungen leicht rückgängig gemacht werden, kann ein System trotzdem in diese Kategorie eingeordnet werden. Systeme, deren Entscheidungen sich schädigend auf Gesundheit und/oder Sicherheit von Personen auswirken können, können grundsätzlich nicht in dieser Kategorie platziert werden.

In die mittlere Kategorie («Hohes Risiko») fallen Systeme,

- die ein hohes Risiko für die Gesellschaft darstellen oder
- die für Individuen
    - ein hohes Risiko für die Gesundheit, Sicherheit oder Grundrechte darstellen.

Die Systeme in dieser Kategorie sind dadurch gekennzeichnet, dass ihrem (positiven) Nutzen ein signifikantes potentielles negatives Potential gegenübersteht. Das Schadenspotential ist dabei noch akzeptabel (oder mitigierbar), ansonsten würde ein solches System in der nächsthöhere Kategorie eingeordnet. Systeme in dieser Kategorie werden typischerweise breit eingesetzt (nicht nur vereinzelt) und lassen den Individuen keine Möglichkeit, sich der automatisierten Entscheidungsfindung zu entziehen. In dieser Kategorie können durch Entscheidungen entstandene Schäden auch realistischerweise nicht rückgängig gemacht oder kompensiert werden.

Entscheidungssysteme, die Inhalte empfehlen (seien es Nachrichten wie bei Newsfeed-Algorithmen, Videos bei Empfehlungsalgorithmen oder allgemeine Inhalte wie bei Suchmaschinen) gehören aus mehreren Gründen in die Kategorie «hohes Risiko»: Sie betreffen grosse Nutzer:innengruppen (potentiell die ganze Gesellschaft), sie beeinflussen die Wahrnehmung ihrer Konsument:innen, und sie können nachgewiesenermassen zur Radikalisierung beitragen (vgl. Tufekci 2018, Frenkel und Kang 2021).

Individuelle Entscheidungen im Sozialwesen (z.B. Berechtigungsbeurteilungen) sind aus mehreren Gründen hochriskant: Sie betreffen in der Regel schutzbedürftige Menschen, können nicht vermieden/umgangen werden und den Betroffenen ist es in der Regel ohne komplizierten und teuren Rechtsweg nicht möglich, Korrekturen von Fehlentscheidungen zu erwirken. Entscheidungen, die zur Einstellung oder Auswahl von Personen in Stellenbewerbungsverfahren führen, haben ebenfalls ein hohes Risiko, da sie von den Betroffenen nicht umgangen werden können, aber die Lebens- und Entwicklungschancen dieser Menschen beeinflussen.

Daneben gibt es Systeme mit dem Potential, sich irreversibel und schwerwiegend auf Individuen auszuwirken, z.B. in der medizinischen Diagnostik, und würden damit als «inakzeptabel» kategorisiert. Solange diese Systeme als Unterstützung eingesetzt werden und die endgültige Entscheidung eine Fachperson trifft, ist eine Rückstufung zu «hohem Risiko» sinnvoll. Jedoch ist der Übergang von Hinweisgeber- über Unterstützungssystem bis zu unhinterfragten Entscheidungen fliessend, besonders in Kombination mit der bereits erwähnten Technologiegläubigkeit, welche anfängliche Unterstützungssysteme zu De-facto-Entscheider mutieren lassen könnte.

In die höchste Kategorie («Inakzeptables Risiko») gehören Systeme,

- die ein inakzeptables Risiko für die Gesellschaft als Ganzes darstellen oder
- die für Individuen
    - ein inakzeptables Risiko für die Gesundheit, Sicherheit oder Grundrechte oder
    - irreversible und schwerwiegende Auswirkungen darstellen.

In diese Kategorie gehören also Systeme, deren potentieller Schaden so gross ist, dass er nicht riskiert werden darf. Bei vielen Systemen in dieser Kategorie ist der Schaden ausserdem bekannt und dokumentiert (und nicht mehr potentiell, so dass man genau genommen nicht mehr von einem Risiko sprechen kann). Die Entscheidungen in dieser Kategorie sind ausserdem weder umgehbar (z.B. biometrische Massenüberwachung) noch revidierbar, sie sind zudem oft auch nicht überprüfbar (z.B. automatisierte/unterstützende Asyl-, Bewährungs- oder Gerichtsentscheidungen). Der erwartete bzw. nachgewiesene Schaden für Individuen und für die Gesellschaft ist in dieser Kategorie so hoch, dass die Risiken nicht akzeptiert und auch nicht mitigiert werden können und der Einsatz solcher Systeme verboten wird.

Beispiele für inakzeptable Auswirkungen für die Gesellschaft sind die bereits oben genannte biometrische Massenüberwachung (inkl. Gesichtserkennung[14]), welche nicht nur einen massiven Eingriff in die Grundrechte, wie Menschenwürde, Autonomie und Privatheit, darstellt, sondern auch mit den erwähnten «Chilling Effects» (vgl. Assion 2014 und Penney 2016) auf die demokratischen Prozesse und damit die Gesellschaft wirkt. Ein weiteres Beispiel ist das automatisierte Bewerten von Verhaltensweisen (Social Scoring), das zwar in erster Linie Individuen betrifft, jedoch durch seine Kontroll- und Formungseffekte weitreichende (und zudem nicht demokratische legitimierte) Auswirkungen auf die Gesellschaft haben kann.

Für Individuen sehen wir neben den Asyl-, Bewährungs- oder Gerichtsentscheiden ebenfalls inakzeptable Auswirkungen bei der Überwachung von Mitarbeiter:innen, Schüler:innen und Student:innen. Weitreichende automatisierte Beurteilung am Arbeitsplatz und daraus folgende Entlassungs- oder Optimierungsentscheide können in inakzeptabler Weise die physische Gesundheit der Arbeitnehmer:innen schädigen.[15]

14 beispielsweise https://gesichtserkennung-stoppen.ch, https://reclaimyourface.eu

15 Aus diesen Gründen wurde bereits gefordert, Überwachung und automatisiertes Management in Arbeits- und Bildungskontexten auf die Liste der zu verbietenden Anwendungen zu setzen (vgl. EDRi 2021); zur Kritik an Techniken der automatisierten Emotionserkennung (vgl. Crawford 2021)

7. Sorgfalts- und Transparenzpflichten

Grundsätzlich ist die aus Sicht der Betroffenen einsetzende Entität des ADM-Systems (zum Beispiel der Dienstleister) für dessen Funktionalität und dessen korrekte Einordnung in die oben genannten Risikokategorien verantwortlich. Zwar soll es möglich sein, gewisse geschäftliche Risiken zivilrechtlich an die Hersteller von Komponenten oder von Systemen weiterzugeben. Jedoch sollte (über die weiter oben erwähnte Produkthaftung des separaten IT-Sicherheits-Gesetzes) verhindert werden, dass sich die Hersteller von jeglicher Verantwortung mittels der AGB entbinden können, so wie dies derzeit in Software-Nutzungsverträgen gängige Praxis ist.

Eine Zertifizierungspflicht erachten wir nur in speziellen und privatwirtschaftlichen Einsatzgebieten mit spezifischem und einfach zu standardisierendem Einsatzzweck[16] wie bei medizinischen Produkten als sinnvoll, z.B. bei einem automatischen Defibrilator (AED). Ansonst besteht die Gefahr, dass die Rechenschaftspflicht grossflächig an Zertifikatsaussteller ausgelagert wird.

Die folgenden Transparenzpflichten gelten nur für die ADM-Systeme der Kategorie «hohes Risiko». «Inakzeptable» Systeme dürfen von vornherein nicht eingesetzt werden. Eine falsche bzw. zu tiefe Deklaration wird mit empfindlichen Strafen geahndet. Der Grad der Transparenzpflichten soll die Einschätzung der einzelnen Systeme bezüglich ihrer Risiken ermöglichen, aber auch ausreichend Informationen zur Abschätzung des Wirkens des gesamten ADM-Ökosystems bieten. Wir empfehlen daher standardisierte Transparenzberichtsformate.

Wir unterscheiden zwischen Pflichten für Systeme, die in der **Privatwirtschaft** eingesetzt werden und solchen, die **in Erfüllung eines öffentlichen Auftrags** verwendet werden. Generell gilt für alle Systeme (privat und öffentlich) mit der Kategorisierung «hohes Risiko» eine **Kennzeichnungs- und Hinweispflicht**, welche

1. darauf hinweist, dass ein ADM-System eingesetzt wird[17],
2. ein kurzes Abstrakt zum Einsatzzweck des Systems und konkrete mögliche Outputs sowie
3. Informationen über die Datenherkunft sowie Erläuterungen zu den spezifischen vom ADM-System genutzten Features (siehe Glossar) und was diese repräsentieren.

Die Informationen zum Datenursprung sollen auch dazu dienen, dass eine ausreichend hohe Datenqualität vorliegt und die **Verkettung** von mehren ADM-Systemen (unter Umständen von verschiedenen Herstellern) besser sichtbar wird. Des Weiteren fordern wir eine periodische und kontinuierliche Überprüfung des Risikos (d.h. der Einstufung der Kategorie) sowie der Do-

16 Dies bedeutet eine klar überprüfbare Funktionsweise des selbständig entscheidenden Systems.
17 mit Anpassung von Art. 21 Abs. 1 nDSG (streichen von «ausschliesslich»)

kumentation bezüglich der Transparenzpflichten, vor allem bei selbständig weiterlernenden Systemen.[18]

Bei der **Datenqualität** geht es darum, dass die verwendeten Daten entweder mit der Realität übereinstimmen,[19] dass darauf basierende Systeme also möglichst fehlerlos funktionieren, oder dass sie nur in beabsichtigen und allgemein nützlichen Aspekten verändert wurden, beispielsweise um Diskriminierung vorzubeugen.

Bezüglich des **Datenherkunft** ist das Recht auf informationelle Selbstbestimmung einzuhalten; dies gilt auch für Daten aus dem Ausland. Daten müssen aus ethisch vertretbaren Quellen stammen, beispielsweise ist von einem Gebrauch illegal beschaffter Daten grundsätzlich abzusehen [20].

Des Weiteren soll konkret ausgewiesen sein, wenn Ausgaben anderer ADM-Systeme verwendet werden. Damit soll Transparenz zur **Verkettung** solcher Systeme geschaffen werden, welche durch die absehbar steigende Komplexität ihres Zusammenwirkens, ihres (opaken) Informationsflusses und der daraus entstehenden Rückkopplungsschleifen eigene Risiken schaffen.

Im Folgenden präzisieren wir die Transparenzpflichten für privatwirtschaftliche und öffentliche Kontexte.

7.1 Privatwirtschaftlicher Kontext

Für Entitäten, welche Systeme innerhalb des privatwirtschaftlichen Kontextes einsetzen, fordern wir Informationen über die Herkunft sämtlicher verwendeter Daten sowie über die Qualität und die Vollständigkeit im Hinblick auf den Zweck des ADM-Systems. Dies inkludiert alle Daten, die zum Aufsetzen, Trainieren, Validieren sowie zur Vorhersage etc. des Systems verwendet werden. Ausserdem umfasst es die Dokumentation zum Zweck des Systems und aussagekräftige Informationen darüber, welche Features als Eingabe verwendet werden, um die Tragweite für Individuen und die Gesellschaft bzw. das Risiko für Gesundheit, Sicherheit oder Grundrechte des Einzelnen oder der Gesellschaft einschätzen zu können.

Ein formelles Gesetz kann in Abwägung von Risiko und Nutzen Ausnahmen zu den Transparenzpflichten sowie der Haftung für standardisierbare Produkte vorsehen, wenn es gleichzeitig eine Zertifizierungsstelle schafft, welche die oben genannten Anforderungen an die Qualität mindestens gleichwertig erfüllt (beispielsweise bei medizinischer Diagnostik).

18 Dies sind Systeme, die Ihre Funktionsweise basierend auf neuen Eingaben kontinuierlich anpassen. Diese Funktionsweise führt zu einer Vielzahl an Problemen, etwa, dass Systeme die zuvor attestierte Garantieren wie «Fairness» verlernen, oder dass sie absichtlich und schwer nachweisbar mit manipulierten Daten gefüttert werden können, um deren Funktionsweise zum eigenen Vorteil zu verändern.

19 Das heisst, sie sind statistisch repräsentativ (bezüglich des Ziels und des Einsatzgebiets des Systems), akkurat, vollständig und möglichst widerspruchsfrei und folgen einer bekannten Semantik.

20 Beziehungsweise ist der Gebrauch nur unter bestimmten Umständen nach einer Abwägung von Vor- und Nachteilen aus ethischer Sicht gerechtfertigt (vgl. Imhasly 2021).

7.2 In Erfüllung eines öffentlichen Auftrags

Neben den selben Transparenzpflichten wie für privatwirtschaftliche Systeme (Informationen über die Datenherkunft und -Qualität, Features und Zweck) fordern wir die Offenlegung der Koeffizienten (siehe Glossar) in standardisiertem Format[21] wie folgt:

- Für ADM-Systeme, die auf Nicht-Personendaten basieren, sollen die Daten als Open-Data soweit möglich zusammen mit den Koeffizienten zur Verfügung gestellt werden.
- Für ADM-Systeme, die auf Personendaten oder auf Nicht-Personendaten basieren, die nicht veröffentlicht werden dürfen, gilt: Sie müssen entweder a) auf synthetisierten Daten (synthetisches Datenset, siehe Glossar) trainiert werden und diese müssen zusammen mit den Koeffizienten veröffentlicht werden; oder b) es werden weder die Daten noch die Koeffizienten veröffentlicht, falls (im Ausnahmefall) die Erzeugung von synthetisierten Daten und die Verwendung entsprechender ADM-Systeme mit unverhältnismässig viel Aufwand verbunden ist oder sich aus deren Koeffizienten Personendaten oder Nicht-Personendaten, die nicht veröffentlicht werden dürfen, wieder ableiten lassen. Jedoch müssen in diesem Fall der ADM-Aufsicht sowie berechtigten NGO Zugang zur Überprüfung der Tragweite für Individuen und für die Gesellschaft bzw. des Risikos für Gesundheit, Sicherheit oder Grundrechte des Einzelnen oder der Gesellschaft ermöglicht werden.

Die hier geforderte generelle Offenlegungspflicht korrespondiert mit der Forderung «Public Money? Public Code!» – der Forderung nach Quell-Code-Veröffentlichung von durch öffentliche Geldern finanzierter Software. Damit können die Lösungen auch von anderen Behörden oder der Öffentlichkeit verwendet und weiterentwickelt werden.

21 Dies macht die automatische Auswertung einfacher.

8. Kontrolle, Massnahmen und Sanktionen

Verletzungen der oben aufgeführten Sorgfalts- und Transparenzpflichten sollen wirksam sanktioniert werden. Auch hier unterscheiden wir zwischen privatwirtschaftlichem und öffentlichem Einsatz. Für beide Fälle wird die Einhaltung der Sorgfalts- und Transparenzpflichten zum einen durch Individuen und zum anderen durch berechtigte Verbände (NGO) kontrolliert, welche im Schadensfall Beschwerde oder Klage führen können. Verbände sollen beschwerdeberechtigt sein, wenn sie gesamtschweizerisch tätig sind und den Zweck in den Statuten verankert haben. Die Kontroll-, die Massnahmen- und die Sanktionsmöglichkeiten sowie die Rechtsmittelwege sind so auszugestalten, dass den Betroffenen der bestmögliche Schutz gewährleistet werden kann; wo nötig, sind diese auch zu ergänzen oder neu auszugestalten. Hierzu gehört auch die Überarbeitung und Verbesserung von kollektiven Rechtsdurchsetzungsmitteln.

Die ADM-Aufsicht soll Verstösse gegen das Gesetz von Amtes wegen untersuchen und formell Verfügungen erlassen können. Sie kann Einsicht verlangen und Sanktionen aussprechen. Um den Betroffenen den bestmöglichen Schutz zu gewährleisten, sollen sowohl vorsätzliches als auch fahrlässiges Handeln strafbar sein. Wir erwarten, dass fragwürdige Systeme schnell bekannt werden, um die Aufmerksamkeit der Zivilgesellschaft und damit der berechtigten Verbände oder der ADM-Aufsicht auf sich zu ziehen.

Falschkategorisierung von ADM-Systemen, beispielsweise als tiefes Risiko anstelle des eigentlich korrekten hohen Risikos, und den damit verbundenen Verstössen gegen die Sorgfalts- und Transparenzpflichten soll vorgebeugt werden, in dem die zu erwartenden Sanktionen hoch genug ausfallen. Die vorgeschlagene Selbstdeklarationspflicht zur Vermeidung von bürokratischen Prozessen und zur Entlastung der Unternehmen sehen wir aus diesem Grund als ausreichend an. Wir erwarten, dass die Unternehmen selbständig Regeln analog zum Datenschutz implementieren, also beispielsweise durch die Einrichtung von internen Meldestellen bei vermuteten Verstössen oder Falschdeklarationen.

8.1 Privatwirtschaft

Der Nachweis einer individuellen Schuld scheint nicht zielführend, handelt es sich doch in der Regel bei Verletzungen der oben aufgeführten Sorgfalts- und Transparenzpflichten um ein Organisationsverschulden. Die ADM-Aufsicht soll daher die Unternehmen mittels Verwaltungssanktionen ahnden und nicht Individuen per Strafrecht sanktionieren. Damit entfällt auch das sonst drohende «Abschieben» der Schuld auf «Sündenböcke». Des Weiteren muss der Strafrahmen umsatzabhängig sein, damit sich Grossunternehmen nicht vergleichsweise günstig aus der Affäre ziehen können. Die Strafen müssen ausreichend hoch sein, damit Verletzungen der Sorgfalts- und Transparenzpflichten nicht als alltägliches Geschäftsrisiko wahrgenommen und somit «mitbudgetiert» werden.

Die zu tiefe Einschätzung der Kategorie des ADM-Systems durch den Verantwortlichen ist strafbar.

Der ADM-Aufsicht stehen insbesondere folgende Instrumente zur Verfügung: Sie sammelt Reklamationen, sie kann Einsicht verlangen und sie kann Sanktionen und Verfügungen erlassen, wie dies beispielsweise bei der FINMA geltende Praxis ist. Die abschliessende Beurteilung obliegt den Gerichten.

Bei vermuteter Unzulänglichkeit sehen wir folgende Einforderungswege:

1. Betroffene Individuen können Klagen gegen Entitäten der Privatwirtschaft und Beschwerde gegen Verfügungen der ADM-Aufsicht führen, falls die Verfügungen der ADM-Aufsicht als unzureichend angesehen werden. Dabei sollen ausdrücklich auch Sammelklagen und -Beschwerden möglich sein. Die Rechtsmittelwege sind in diesem Sinne anzupassen.
2. Berechtigte Verbände (gesamtschweizerisch und mit passendem Zweck gemäss Statuten) sollen ohne persönliche Betroffenheit eine Klage gegen private Entitäten und Beschwerden gegen Verfügungen der ADM-Aufsicht führen können (Verbandsbeschwerde- resp. -klagerecht). In Anbetracht der hohen Prozessführungskosten und als regulatives Element kann der entsprechende Verband einen Teil des Sanktionsbetrags als Aufwandsentschädigung erhalten.

Im Falle einer drohenden Verurteilung führt das Verschleierungsinteresse der Angeklagten zu einem starken Machtungleichgewicht. Wir fordern daher die Beweislastumkehr, so dass beschuldigte Entitäten hinreichend belegen müssen, dass sie die Kategorisierungsvorgaben, Sorgfalts- oder Transparenzpflichten nicht verletzt haben, falls ein Gericht die Beschwerde oder Klage als zulässig anerkennt.

8.2 In Erfüllung eines öffentliches Auftrags

Grundsätzlich sollen die gleichen Kontrollen, Massnahmen und Sanktionen wie gegen private Entitäten möglich sein. Wie das Verhältnis zwischen der ADM-Aufsicht und den Entitäten in Erfüllung eines öffentlichen Auftrags auf kantonaler und kommunaler Ebene im Detail auszugestalten ist, bleibt zu klären.

Es soll sowohl für Individuen als auch für Verbände Möglichkeiten geben (analog Kapitel 8.1 Privatwirtschaft), sowohl gegen Risiken, die von ADM-Systemen ausgehen, als auch gegen Ergebnisse derartiger Systeme vorzugehen. Um allfällige Kompetenzkonflikte zu umgehen, könnten beispielsweise, wenn Entitäten im öffentlichen Auftrag auf kommunaler oder kantonaler Ebene betroffen sind, der ADM-Aufsicht im kantonalen Verfahren immer Parteirechte zukommen.

9. Einige Anregungen für die Zukunft

Systeme für automatisierte Entscheidungen werden zukünftig immer mehr Aufgaben, Arbeiten und Funktionen übernehmen, woraus sich neue Konsequenzen, Chancen, Herausforderungen und Probleme ergeben können. Im Folgenden wollen wir daher im Sinne einer Technologiefolgeabschätzung verschiedene Punkte ansprechen, für die ein regulatorisches Eingreifen nötig werden könnte.

Der erste Punkt betrifft die **Machtfrage**: Wer erschafft, bestimmt und kontrolliert die eingesetzten Systeme, Algorithmen und Metriken? Die entsprechenden Personen und Organisationen haben starken Einfluss auf die Wahrnehmung und auf Möglichkeiten unserer natürlichen und sozialen Umwelt. Hier gilt es entsprechend sehr genau hinzuschauen, wie sich diese Abhängigkeiten entwickeln.

Der zweite betrifft die **Vernetzung und Verkettung** von weitreichenden automatisierten Systemen: In naher Zukunft könnte die Ausgabe eines Systems teilweise die Eingabe des anderen Systems sein, welches wiederum einen Einfluss auf das erste System haben kann. Dies kann, vor allem mit mehr als nur zwei Systemen, zu komplexen und vielschichtigen Rückkopplungseffekten (siehe Anhang) und damit schwer abzuschätzenden Risiken führen. Die absehbare, partielle Intransparenz der verketteten Systeme und die damit einhergehende Unvorhersehbarkeit dieser Effekte wird eine Auseinandersetzung damit nötig machen. Potentielle Lösungsansätze wären eine klare Modularisierung der Systeme, sodass das interne Wirken der Systeme auf eine einfache Abstraktion reduziert werden kann, und dass dies ausreicht, um die Folgen der Rückkopplung abzuschätzen. Denkbar ist auch ein Kopplungsverbot für Systeme ab einer gewissen Cluster-Grösse oder wenn gewisse Sicherheits- oder Zweckbindungskriterien nicht mehr erfüllt sind.

In gewissen Kreisen besteht die Vision, dass sich alle soziale und persönlichen Probleme mit mehr Daten und besseren Algorithmen lösen lassen, wenn man sie nur zulässt. Diese Weltsicht versucht, die komplette Realität in ein Konstrukt aus Formeln und Zahlen zu pressen. Aufgrund unserer Einsicht, dass es keine absolute Objektivität gibt und dass daher alle Metriken, Mess- und Kennzahlen sowie deren Interpretation Gegenstand gesellschaftlicher Aushandlungsprozesse sind, sehen wir diesen Weg als irreführend an. Wir raten daher zur generellen **Datensparsamkeit als Grundprinzip**, da diese, wie beim Datenschutz auch, die entstehenden Probleme bereits an der Quelle reduziert.

Weiter ist eine **Abhängigkeit von ADM-Systemen** absehbar. Der Einsatz von Automation zur Arbeitserleichterung und -abnahme ermöglicht, mehr und komplexere Aufgaben in kürzerer Zeit zu bewältigen. Wir sollten uns aber bewusst sein, was etwa ein Ausfall dieser automatisierten Systeme für uns bedeuten würde, welche Reichweite er hätte und welche Risiken damit verbunden wären, und als Massnahme schnell umsetzbare Notfallstrategien vorbereiten. Durch die zunehmende Vernetzung und die Abhängigkeit von einzelnen Ressourcen, wie etwa im Falle des Internets, steigt auch die Gefahr, dass viele Funktionen gleichzeitig ausfallen

könnten. Vielleicht macht es Sinn, über Massnahmen zu sprechen, die komplett redundante Systeme hervorbringen.

Im selben Zug kann man auch einen potentiellen **Kompetenzverlust des Menschen** und einen **Verlust von Verantwortlichkeit** erahnen. Die Abgabe von Aufgaben an automatisierte Systeme, das Sich-Verlassen auf die korrekte Ausführung und die damit verbundenen Habituationseffekte könnten zu einem Verlernen von Kompetenzen, die ohne entsprechende Systeme benötigt würden, und zu einer geringeren individuellen oder kollektiven Verantwortlichkeit führen. Entsprechende Effekte könnten sich womöglich erst im Verlaufe von mehreren Generationen zeigen, etwa wenn gewisse Fähigkeiten nicht mehr weitergegeben werden.

Einfache Jobs ohne lange Ausbildungsanforderungen werden immer mehr verloren gehen. Dies kann erhebliche soziale Folgen, auch in der Schweiz, mit sich bringen, da unter anderem sozialer Status und Arbeit eng verknüpft sind. Wir müssen über unser Verständnis von sozialer Wertschätzung und damit auch über die Verteilung von Wohlstand nachdenken und diese womöglich langfristig umdefinieren, damit **alle Menschen am Gewinn der Automatisierung teilhaben** können.

Schliesslich möchten wir die Bedeutung einer kontinuierlichen **Technologiefolgeabschätzungen** hervorheben, wie dies beispielsweise die TA-Swiss im Auftrag des Bundes neben anderen Organisationen macht. Das generelle Ziel der Technologiefolgeabschätzung ist, eine systematische Analyse und Bewertung der Auswirkungen und Folgen von Technologien in allen ersichtlich betroffenen Teilbereichen der natürlichen und sozialen Umwelt zu erstellen.

Anhang

A. Feedback-Loops

Feedback-Loops (Rückkopplungsschleifen) beschreiben im ADM-Umfeld die Auswirkung von Resultaten von AMD-Systemen auf ihren Input oder auf den Input ähnlich agierender Systeme. Diese Effekte sind vielfältig und können sich über mehrere ADM-Systeme und ihr Wechselwirken mit dem (reaktiven) Verhalten der Menschen erstrecken. Aufgrund dieser komplexen Interaktionen und Kreisläufe sind diese Rückkopplungsschleifen schwer zu erkennen und oft erst nach einiger Zeit indirekt sichtbar. Besonders problematisch werden sie, wenn sich schädliche Dynamiken selbst verstärken.

Ensign et al. (2018) beschreiben dazu eine Klasse solcher Rückkopplungsschleifen am Beispiel von Predictive-Policing-Systeme (vgl. Ensign 2018). Sie belegen (mathematisch und empirisch), dass die Vorhersage von bevorstehenden Verbrechen in bestimmten Regionen zu mehr Polizeipräsenz in diesen Regionen geführt hat. Dadurch wurde dort eine verhältnismässig höhere Anzahl an Verbrechen als in Regionen mit weniger Polizeipräsenz detektiert, was dann zu noch mehr Polizeipräsenz in den vorbelasteten Einsatz-Regionen geführt hat. Diese Dynamik verstärkt sich selber.

Das Grundproblem dieser Art von «Runaway»-Feedback-Loops liegt darin, dass nur eine Art von Resultaten wieder Eingang in die Systeme finden (die detektierten Verbrechen) und andere ausgeblendet werden (die nicht-detektierten Verbrechen in anderen Regionen). Dies führt zu statistischen Verzerrungen der Ausgangsdatenlage, welche die Trainingsbasis für den Predictive-Policing-Algorithmus bildet. Ähnliches trifft zu bei Systemen zur Selektion von Bewerber:innen auf eine Arbeitsstelle, wo bis anhin meist nur die Personen mit erhaltener Anstellungen berücksichtigt wurden. Ein weiteres Beispiel ist die numerische Bewertung der Universitäten in den USA, die zu einem teuren «Wettrüsten» bezüglich der angebotenen Dienstleistungen geführt hat und aufgrund des dadurch erhöhten Kapitalbedarfs in einer starken Zunahme der Semestergebühren resultierte (vgl. O'Neil 2016).

Eine andere Art von Feedback-Loop illustriert Cathy O'Neil (2016) anhand des Lehrpersonenbewertungssystems IMPACT. IMPACT versuchte die Lehrleistung der Lehrer:innen als Differenz zwischen der vorhergesagten «natürlichen» Entwicklung der Schüler:innen basierend auf ihrer Herkunft und ihrer bisheriger Leistung und durch von ihren Schüler:innen geschriebenen Tests zu eruieren, um auf dieser Basis regelmässige Entlassungen von Lehrpersonen vorzunehmen. Dies ist in sich schon problematisch, da sich das komplexe soziale Umfeld heranwachsender Schüler:innen nicht nur in ein paar Zahlen ausdrücken lässt und da 20 oder 30 Schüler deutlich zu wenig sind, um für solche Modelle statistisch belastbare Resultate zu erhalten. Daher korrelierte die Ausgaben dieses Algorithmus zur selben Lehrperson über mehrere Jahre auch nicht stark, was der beabsichtigten Funktion des Algorithmus (das objektive Bewerten von Lehrpersonen) widersprach. Zusätzlich haben einige Lehrpersonen begonnen, ihre Schüler:innen und die Tests so zu manipulieren, dass für sie und ihre Zukunft als angestellte Lehr-

personen vorteilhafte Ergebnisse entstanden, was sich auch zum Nachteil jener auswirkte, die das nicht taten. Diese Art von destruktiver sozialer Feedback-Loop konnte der Algorithmus nicht korrigieren. Das Projekt wurde aufgegeben.

Hauptsächlich aufgrund ihres begrenzten Einsatz- und Wirkungsbereiches konnten die erwähnten Beispiele der Verbrechensvorhersage oder der Lehrpersonenbewertung durch menschliche Einsicht (nicht automatisiert) und entsprechend angelegte Randbedingungen korrigiert oder als Projekt komplett aufgeben werden. Vergleichbare Effekte können aber auch über deutlich kompliziertere Verknüpfungen von Systemen und Gesellschaft auftreten, wo eine Korrektur durch einzelne Parteien nicht mehr möglich oder eine offensichtliche Lösung nicht ersichtlich ist. Bei Predictive-Policing zum Beispiel handelt es sich zwar um ein Online-Learning-System (das heisst um ein kontinuierlich dazulernendes System), doch das Problem entsteht auch, wenn regelmässig neue ADM-Systeme basierend auf neu gesammelten Datensätzen erschaffen werden. In diesem Fall wird die Feedback-Loop über den gesellschaftlichen Bias (Verzerrung) in den Trainingsdaten (siehe Glossar) konstruiert. Es reicht daher nicht, für diesen Effekt nur Online-Learning-Systeme zu berücksichtigen.

B. Regulierungsvorschläge für ADMS, Künstliche Intelligenz und Algorithmen

Im Zuge der Digitalisierung haben sich (datengetriebene) Informatiksysteme praktisch unreguliert ausgebreitet (abgesehen von hauptsächlich europäischen Datenschutzgesetzen). Je nach Aspekt, der betont werden soll, werden diese System als «Automatisierte Entscheidungssysteme», «Algorithmische Systeme», «Künstliche Intelligenz-», «Artificial Intelligence-» oder «Big-Data-Systeme» bezeichnet. Bei allen Unterschieden haben diese gemeinsam, dass sehr grosse Datenmengen verarbeitet und analysiert werden mit dem Ziel der Automatisierung von Entscheidungen und/oder Prozessen. In den letzten Jahren hat sich ein breiter Konsens gebildet, dass diese Systeme reguliert werden müssen, um die stärksten negativen Auswirkungen und Risiken zu vermeiden. Die Forderung nach Regulierung kommt dabei nicht nur aus der Zivilgesellschaft, sondern auch aus Politik, Wirtschaft und Forschung.

So hat zum Beispiel die ACM (Association for Computing Machinery) bereits 2017 eine Positionierung zur Transparenz und Verantwortlichkeit von Algorithmen erarbeitet (vgl. ACM 2017). Die ACM ist der Berufsverband der US-amerikanischen Informatiker:innen und somit die Organisation, der die Menschen angehören, die Forschung, Entwicklung und Einsatz von ADM-System an vorderster Stelle vorantreiben. Viele der Aussagen und Prinzipien in dieser Positionierung, beispielsweise bezüglich Transparenz und Daten, finden sich auch in unserem Vorschlag wieder.

Auch aus der Wirtschaft kommen immer wieder Aufforderungen an die Politik, solche Systeme zu regulieren. Besonders eindrücklich ist das Statement des Microsoft-Präsidenten Brad Smith von 2018, in dem er betont, dass Gesichtserkennung aufgrund seines dystopischen Potentials und seiner Gefahren für die Demokratie reguliert werden müsse (vgl. Smith 2018).

Bisherige Ansätze und Vorschläge (vgl. DEK 2019, EU AI Act 2021) verfolgen einen risikobasierten Ansatz, in dem versucht wird – wie auch unser Vorschlag dies tut –, ADM-Systeme anhand ihres immanenten Risikos in Kategorien einzuteilen und für die Kategorien mit steigendem Risiko strengere Regeln zu definieren. Die Anzahl der Kategorien variiert dabei zwischen den Vorschlägen. Es gibt jedoch immer eine (risikolose bzw. -arme) tiefste Kategorie mit sehr wenigen Regeln bzw. Anforderungen sowie eine als sehr riskant eingestufte Kategorie von ADM- Systemen, deren Einsatz verboten wird. Im Vorschlag der Datenethikkommission wurde so beispielsweise die Risikopyramide entwickelt.

Auch der Vorschlag der EU-Kommission «AI Act» (vgl. EU AI Act 2021), der im April 2021 veröffentlicht wurde, folgt diesem risikobasierten Ansatz. Im Unterschied zu unserem Vorschlag enthält der AI ACT konkrete Aufzählungen von verbotenen (wie «remote post biometric identification») und hochriskanten Anwendungen. Seit seiner Veröffentlichung wird der AI Act intensiv diskutiert. Während eine breite Übereinstimmung existiert, was die Notwendigkeit und Relevanz dieses Vorschlags sowie seinen allgemeinen, risikobasierten Ansatz anbelangt, gibt es darüber hinaus auch detaillierte Kritik aus der Zivilgesellschaft (so fordert die Digitale Gesellschaft mit einer grossen Allianz des EDRi-Netzwerks unter anderem eine breitere Fassung der verbotenen und hochriskanten Kategorien beziehungsweise eine Streichung der Ausnahmen, insbesondere im Bereich der biometrischen Identifikation, vgl. EDRi 2021). Auch von Konsumentenschutzorganisationen kommt ähnliche Kritik (vgl. VZBV 2021). Frankreich führte schon 2016 eine Veröffentlichungspflicht für ADM-Systeme öffentlicher Behörden ein, die für Individuen automatisierte administrative Entscheide treffen (vgl. Journal officiel 2016).

Für die Schweiz gibt es bereits ein Positionspapier zur Regulierung der KI (vgl. Thouvenin et al. 2021). Wie unser Vorschlag auch, betont dieses Papier die Notwendigkeit einer Regulierung. Im Unterschied zum Vorschlag der Digitalen Gesellschaft lässt es jedoch weitgehend offen, wie diese Regulierung aussehen könnte.

Das AI Now Institute hat eine ganze Reihe von staatlichen «Use Cases» für die Stadt New York zusammengestellt (vgl. AI Now Institute 2018), von denen viele auch auf die Schweiz übertragbar sind. Der Bericht enthält zudem weiterführende Referenzen und motivierende Beispiele.

Quellenverzeichnis

Quellen hinsichtlich Regulierungsvorschläge von ADM-Systemen im europäischen sowie interkontinentalen Kontext

- ACM U.S. Public Policy Council (2017): Statement on Algorithmic Transparency and Accountability
- CAHAI - Ad hoc Committee on Artificial Intelligence of the Council of Europe (2021): A legal framework for AI system (https://edoc.coe.int/en/artificial-intelligence/9648-a-legal-framework-for-ai-systems.html)
- Datenethikkommission der Bundesregierung (2019): Gutachten der Datenethikkommission der Bundesregierung. Berlin: Bundesministerium des Innern, für Bau und Heimat. Zugriff über https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.html (Zugriff am 16.12.2019).
- EU AI Act (2021): Vorschlag für eine Verordnung des europäischen Parlaments und des Rates COM/2021/206 vom 21. April 2021 zur Festlegung harmonisierter Vorschriften für Künstliche Intelligent (Gesetz über künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union
- EU Digital Service Act (2020): Vorschlag für eine Verordnung des europäischen Parlaments und des Rates COM/2020/825 vom 15. Dezember 2020 über einen Binnenmarkt für digitale Dienste (Gesetz über digitale Dienste) und zur Änderung der Richtlinie
- European Commission adoption consultation (2021): Artificial Intelligence Act. EDRi (https://edri.org/wp-content/uploads/2021/08/European-Digital-Rights-EDRi-submission-to-European-Commission-adoption-consultation-on-the-Artificial-Intelligence-Act-August-2021.pdf)
- Europan Union Agency for Fundamental Rights (2019): Facial recognition technology: fundamental rights considerations in the context of law enforcement. Europan Union Agency for Fundamental Rights, November 2019 (https://fra.europa.eu/en/news/2019/facial-recognition-technology-fundamental-rights-considerations-law-enforcement), (Zugriff am 08.02.2022)
- Florent Thouvenin, et al (2021): Ein Rechtsrahmen für Künstliche Intelligenz, Digital Society Initiative Universität Zürich
- Journal officiel (2016): Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique, Journal officiel de la Republique francaise du 8 octobre 2016
- Richardson, Rashida et al. (2019): Confronting Black Boxes: A Shadow Report of the New York City Automated Decision System Task Force. AI Now Institute ainowinstitute.org/ads-shadowreport-2019.html

- Verbraucherzentrale Bundesverband (2021): Artificial Intelligence needs Real-World Regulation, https://www.vzbv.de/sites/default/files/2021-08/21-08-03_vzbv_Position_Paper_AIA_ENG.pdf

Weitere Quellen

- AI Now Insititute (2018): Automated Decision Systems - Examples of Government Use Cases, https://ainowinstitute.org/nycadschart.pdf (Zugriff am 08.02.2022)
- Assion, Simon (2014): Überwachung und Chilling Effect, in: «Überwachung und Recht», Tagungsband zur Telemediucs Sommerkonferenz 2014, epubli GmbH, Berlin
- Bürgi, Urs (2022): «Arbeitnehmerüberwachung», in: Law Media, 2022, https://www.arbeits-recht.ch/fuersorgepflicht-datenschutz/arbeitnehmerueberwachung (Zugriff am 25.01.2022), Urheber: Bürgi Nägeli Rechtsanwälte
- Crawford, Kate (2021): Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press, New Haven, MA
- Ensign, Danielle, et al. (2018): Runaway Feedback Loops in Predictive Policing. FAT 2018: 160-171
- Frenkel, Sheera und Kang, Cecilia (2021): An Ugly Truth. Harper Collins Publishers
- Gesichtserkennung-stoppen.ch (2021): https://www.gesichtserkennung-stoppen.ch/ (Zugriff am 14.02.2022)
- Imhasly, Patrick (2021): Artikel Forschung an Raubgut, in: NZZ am Sonntag vom 19.09.2021
- O'Neil, Cathy (2016): Weapons of Math Destruction. New York: Crown
- Penney, Jonathon (2016): Chilling Effects: Online Surveillance and Wikipedia Use, in: Berkeley Technology Law Journal, Vol. 31, No. 1, p. 117, 2016, Available at SSRN: https://ssrn.com/abstract=2769645
- Public Code, Public Money: https://publiccode.eu/ (Zugriff am 14.02.2022)
- Reclaim your Face (2021): https://reclaimyourface.eu/ (Zugriff am 14.02.2022)
- Smith, Brad (2018): Facial recognition technology: The need for public regulation and corporate responsibility, In: blogs.microsoft.com (https://blogs.microsoft.com/on-the-issues/2018/07/13/facial-recognition-technology-the-need-for-public-regulation-and-corporate-responsibility/) (Zugriff am 08.02.2022)
- Tufekci, Zeynep (2018): YouTube, the Great Radicalizer, in: The New York Times, 10.3.2018, https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html (Zugriff am 8.2.2022)
- Willson, Michele (2017): Algorithms (and the) everyday, in: Information, Communication & Society, 20:1, 137-150, DOI: 10.1080/1369118X.2016.1200645

Glossar

- **ADM-Systeme**: Siehe automatisierte Entscheidungssysteme.
- **Algorithmus**: Ein Algorithmus ist eine eindeutige und schrittweise Handlungsvorschrift zur Lösung eines Problems oder einer Klasse von Problemen, welche nach einer endlichen Anzahl an Rechenschritten zur Lösung kommt. Mittels Stift und Papier können Menschen auch Algorithmen ausführen.
- **Automated Decision-Making Systeme**: Siehe automatisierte Entscheidungssysteme.
- **Automatisierte Entscheidungssysteme**: (Englisch: Automated Decision-Making Systems). Dies ist jede Software, jedes System oder jeder Prozess, der darauf zielt, menschliche Entscheidungsfindungen zu automatisieren, zu unterstützen oder zu ersetzen. Automatisierte Entscheidungssysteme können zum einen aus Werkzeugen zum Analysieren von Datensets bestehen, welche (numerische) Bewertungen, Vorhersagen, Klassifikationen oder Handlungsempfehlungen erstellen. Sie können zum Fällen von Entscheidungen benutzt werden, die einen Einfluss auf das Wohlergehen von Menschen haben. Dieses Wohlergehen umfasst (nicht abschliessend) Entscheide zu sensiblen Lebensbereichen wie Ausbildungsmöglichkeiten, Gesundheitsergebnisse, Arbeitsleistung, Job-Möglichkeiten, Mobilität, Interessen, Verhalten und persönliche Autonomie. Zum anderen können unter automatisierten Entscheidungssystemen auch die Prozesse verstanden werden, welche derartige Werkzeuge implementieren. (nach AI Now, Richardson et al. 2019, S. 20, unsere Übers.)
- **Bias**: Auch: Verzerrung oder Vorurteil. Algorithmische Bias treten auf, wenn ein Computersystem die impliziten Werte der Menschen widerspiegelt, die am Kodieren, Sammeln, Auswählen oder Verwenden von Daten zum Trainieren des Algorithmus beteiligt sind.
- **Daten**: Auch Datenset. Eine Sammlung von Datenreihen, welche für das Aufsetzen, Trainieren, Validieren, Vorhersage etc. von ADM-Systemen verwendet wird.
- **Datenreihe**: Eine Kollektion von Zahlen, Texten, Bildern, Graphen und so weiter (für Computer sind das alles Zahlen), die sich auf eine einzelne Person, ein bestimmtes Ereignis oder einen gemessenen Umstand beziehen.
- **Features**: Features sind Datenattribute der Datenreihen und davon abgeleiteten Datenattribute, die als Eingangsdatenreihen für den Entscheidungsalgorithmus (das Modell) verwendet werden. Dies können zum Beispiel Alter, Postleitzahl, Mineralwasservorliebe, aber auch davon abgeleitete Meta-Variablen wie Ernährungsgesundheit sein.
- **Feedback-Loop**: Feedback-Loops (Rückkopplungsschleifen) beschreiben im ADM-Umfeld die Auswirkung von Resultaten von ADM-Systemen auf ihren Input oder auf den Input ähnlich agierender Systeme. Für detaillierte Erläuterungen siehe Anhang Kapitel A.
- **Koeffizienten**: Koeffizienten sind Konkrete Zahlen, welche nach der Verrechnungsvorschrift (der Architektur des Modells) zusammen mit den Eingangsdatenreihen verrech-

net werden, um eine Vorhersage durch das Modell zu erhalten. Diese sind beispielsweise die Gewichte bei Neuronalen Netzen oder die Unterscheidungsgrenzen bei Entscheidungsbaum-Algorithmen (Decision-Trees).

- **Modell (Entscheidungsalgorithmen):** Ein Algorithmus (siehe Algorithmus), welche in Eingangsdatenreihen gewisse Typen von Mustern und Zusammenhängen erkennen kann. Dabei werden die Zahlen der Eingangsdatenreihen mit anderen Zahlen (die Koeffizienten des Modells) nach der Verrechnungsvorschrift (der Architektur des Modells) verrechnet.
- **Rückkopplungsschleife**: Siehe Feedback-Loop.
- **Synthetisches Datenset**: Künstlich erzeugte Datenreihen, die in allen wesentlichen Merkmalen echten Datenreihen entsprechen. Der Einsatz von synthetischen Daten vermeidet datenschutzrechtliche Probleme beim Einsatz von sensitiven Daten wie Personendaten. Synthetische Datensets werden zwar künstlich erzeugt und deren einzelne Datenreihen können keiner realen Person oder keinem realen Objekt zugeordnet werden. Aber sie können die Eigenschaften, die ein spezifischer Algorithmus darauf vorhersagen will, korrekt abbilden, sodass Algorithmen, die auf diesen synthetischen Daten trainiert werden, auch auf realen Datenreihen die entsprechende Eigenschaft korrekt ableiten können. Einfach gesagt besitzen synthetische Datensets dieselben relevanten Eigenschaften wie reale Datensets, sodass man Algorithmen darauf trainieren kann, die auf synthetischen sowie realen Datenreihen funktionieren. Sobald jedoch Eigenschaften aus synthetischen Datensets abgeleitet werden sollen, die bei ihrer Erstellung aus dem realen Datenset nicht berücksichtigt wurden, kann dies fehlschlagen.
- **Trainingsdaten**: Eine Kollektion von Datenreihen, die für die Entwicklung respektive das Training von ADM-System eingesetzt werden.
- **Validierungsdaten**: Eine Kollektion von Datenreihen (typischerweise unabhängig von den Trainingsdaten), um die Genauigkeit eines trainiertes ADM-System zu evaluieren.