September 2024
Version 2.0

DIGITALE GESELLSCHAFT

POSITIONSPAPIER

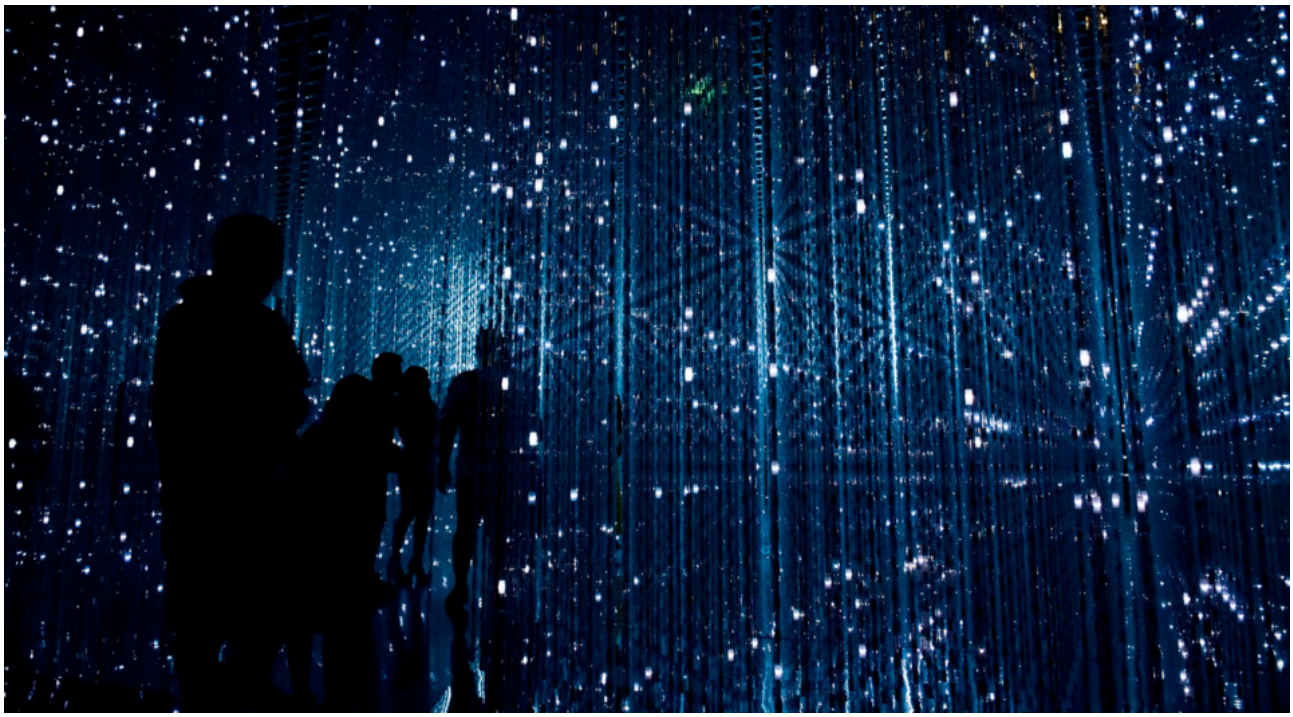
ZUR REGULIERUNG VON AUTOMATISIERTEN ENTSCHEIDUNGSSYSTEMEN

Mitarbeitende der Version 2.0:

David Sommer, Lars Lünenburger, Erik Schönberger, Andreas Geppert, Luka Nenadic, Christian Sigg, Tanja Klankert, Peter Fröhlich, David Caspar, Thomas Mandelz und Jan Zwicky

Mitarbeitende der Version 1.0:

David Sommer, Andreas Geppert, Christian Sigg, David Caspar, Erik Schönenberger, Jan Zwicky, Lars Lünenburger, Luka Nenadic, Peter Fröhlich, Tanja Klankert und Thomas Mandelz



DIGITALE GESELLSCHAFT / CH-4000 BASEL
OFFICE@DIGITALE-GESELLSCHAFT.CH / DIGITALE-GESELLSCHAFT.CH

Executive Summary

Automatisierte Entscheidungssysteme (ADM-Systeme, ADMS) bieten ein grosses gesellschaftliches und wirtschaftliches Potenzial. Sie bergen aber gleichzeitig erhebliche Risiken für Individuen und die Gesellschaft, vor welchen die aktuelle Gesetzgebung in der Schweiz keinen genügenden Schutz bietet. Die Digitale Gesellschaft stellt deshalb ihren Vorschlag für einen Rechtsrahmen zur Regulierung von ADM-Systemen vor, der auf folgenden fünf Pfeilern beruht:

- Beim Einsatz von ADM-Systemen ist der Schutz der Individuen und der Gesellschaft sicherzustellen. Die grundsätzlichen **Schutzziele** in Bezug auf Individuen sind die Einhaltung der Grund- und Menschenrechte. Wenn viele Individuen, gesellschaftsweite Prozesse oder demokratische Organisationen betroffen sind, sprechen wir von Schutzzielen in Bezug auf die Gesellschaft. Die Schutzziele orientieren sich am Datenschutz-Konzept der Digitalen Gesellschaft.
- ADM-Systeme werden in **drei Risikokategorien** eingeteilt: ADM-Systeme haben ein *«tiefes Risiko»*, wenn sie aller Voraussicht nach keine besonderen negativen Auswirkungen auf Individuen oder die Gesellschaft haben. Ein ADM-System mit *«hohem Risiko»* – etwa beim Einsatz in heiklen Bereichen wie dem Sozialwesen – birgt ein signifikantes Schadenspotenzial für die Individuen oder für die Gesellschaft. Schliesslich werden ADM-Systeme mit *«inakzeptablem Risiko»* wie die biometrische Massenüberwachung gänzlich verboten. Um unnötige Bürokratie zu vermeiden, erfolgt die Einteilung auf der Basis von *Selbstdeklarationen*. Deren Korrektheit wird durch nachträgliche Sanktionen bei Falschdeklarationen sichergestellt.
- Um Innovationen nicht zu hemmen, sollen Unternehmen nicht mit übermässigen ex ante Pflichten überhäuft werden. Dieser Gestaltungsspielraum soll aber gleichzeitig nicht als Freipass für einen verantwortungslosen Einsatz von ADM-Systemen missbraucht werden. Es ist zentral, dass Unternehmen und Behörden für Schäden, die durch den Einsatz von ADM-Systemen entstehen, vollständig einstehen. Im Ergebnis verfolgt der Vorschlag daher eine **Mischform zwischen risiko- und schadensbasiertem Ansatz.**
- Im Gegenzug zu diesen Gestaltungsfreiheiten sollen für die Kategorie «hohes Risiko» bestimmte **Sorgfalts- und Transparenzpflichten** gelten, insbesondere in Bezug auf Datenqualität und -herkunft. Durch die öffentliche Hand eingesetzte ADM-Systeme sollen noch strengere Voraussetzungen erfüllen.
- Es werden **griffige Aufsichts- und Sanktionsmechanismen** benötigt. Die Digitale Gesellschaft fordert insbesondere eine staatliche ADMS-Aufsicht, abschreckende Strafen gegen Unternehmen und kollektive Rechtsbehelfe für die Betroffenen.

Inhaltsverzeichnis

1 Einleitung 3

2 Geltungsbereich 4

3 Zusammenfassung Rechtsrahmen 5

4 Die gesellschaftliche Relevanz 6

5 Ein Regulierungsvorschlag für ADM-Systeme 7
5.1 Die ADMS-Aufsicht 9
5.2 Die IT-Sicherheit 9

6 Kategorisierung 10
6.1 Schutzziele und Risiken für Individuen und die Gesellschaft als Ganzes . . . 10
6.2 Beurteilungskriterien 11
6.3 Die Kategorien 12

7 Sorgfalts- und Transparenzpflichten 14
7.1 Privatwirtschaftlicher Kontext 15
7.2 In Erfüllung eines öffentlichen Auftrags 15

8 Kontrolle, Massnahmen und Sanktionen 16
8.1 Privatwirtschaft 16
8.2 In Erfüllung eines öffentlichen Auftrags 17

9 Einige Anregungen für die Zukunft 17

A Regulierungsvorschläge für ADM-Systeme, Künstliche Intelligenz und Algorithmen 19

B Quellenverzeichnis 20
B.1 Quellen hinsichtlich Regulierungsvorschläge von ADM-Systemen im europäischen sowie interkontinentalen Kontext 20
B.2 Weitere Quellen 20
B.3 Bildnachweis 21

C Glossar 21

D Änderungstabelle 23

1 Einleitung

Künstliche Intelligenz (KI) oder Systeme für automatisierte Entscheidungen (Automated Decision-Making Systems, ADM-System, ADMS, siehe Glossar) sind kein futuristisches Wunschdenken mehr. Sie sind in unserem Alltag bereits im Einsatz, unterstützen uns in der Entscheidfindung, vereinfachen die Interaktion mit Computern und können komplexe Texte, Bilder oder Musik generieren. Das positive Potential ist immens. So werden wir beispielsweise mit Maschinen sprechen können, mühselige Arbeiten werden weiter automatisiert, und die personalisierte Medizin revolutioniert die bisherigen Methoden der Diagnose sowie Medikation. Die für Medizinprodukte zuständige amerikanische Zulassungsbehörde FDA listet bereits einige Hundert Medizinprodukte, die KI und Machine Learning verwenden (Stand 13.05.2024 sind es 882).[1]

Doch wie bei den meisten technologischen Umwälzungen zeigen sich auch negative Seiten, beim Thema KI und ADM-Systemen nicht zu knapp. Der wohl bekannteste Fall ist die Kindergeldaffäre in den Niederlanden, bei der zu Unrecht Kindergeld von Tausenden von Familien zurückgefordert wurde, da beispielsweise mehrfache Staatsangehörigkeit als Indikator angesehen wurde.[2] Es ist heute gar nicht mehr möglich, abschliessend zu beurteilen, in welchen Bereichen Künstliche Intelligenz eingesetzt wird und in welchen nicht. Und wir stehen erst am Anfang einer Veränderung, deren Ausmass wir nicht abschätzen können. Es gilt, diese Entwicklung derart zu steuern, dass wir von der positiven Seite profitieren und gleichzeitig die negativen Auswirkungen minimieren können.

Andere Staaten und Staatenverbunde – darunter die EU, China oder die USA – haben das transformative Potenzial von KI und ADM-Systemen erkannt. Sie versuchen aktiv, solche Systeme in rechtlich geregelte Bahnen zu lenken. Die Digitale Gesellschaft hat sich aktiv in diese Debatten eingebracht[3] und fordert auch in der Schweiz eine entsprechende Anpassung der bestehenden Regularien an die verschobenen Gleichgewichte und neuen Herausforde-

[1] Siehe Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices | FDA

[2] Siehe den Artikel: Aufsicht und Transparenz: Wie die Niederlande aus KI-Skandalen lernen (netzpolitik.org)

[3] Dies zum Beispiel bei der Rahmenkonvention zu KI, Menschenrechten, Demokratie und Rechtsstaatlichkeit des Europarats, siehe das Dossier ADM-Systeme der Digitalen Gesellschaft für mehr Informationen.

rungen durch KI- und ADM-Systeme, sodass Nutzen und Risiken in einem guten Verhältnis zueinander stehen.

In diesem Dokument führen wir unseren Vorschlag eines Rechtsrahmens für die Schweiz aus. Der Vorschlag ist technologieneutral[4] und folgt einem «human-centered» Ansatz: Die Systeme sollen dem Menschen nutzen, das heisst es soll dem Menschen durch den Einsatz von KI und ADM-Systemen besser gehen.

Der Einsatz von ADM-Systemen ist an die Bedingungen *der Transparenz* und *der Nachvollziehbarkeit* zu knüpfen. Ein ADM-System erfüllt die Bedingung der *Transparenz*, wenn die Grundlagen der automatisierten Entscheidfindung offengelegt sind. Hat beispielsweise ein ADM-System, welches zur Prüfung der Reintegration von verunfallten Personen in den Arbeitsmarkt eingesetzt wird, alle relevanten Fakten des konkreten Sachverhalts erfasst und gewürdigt? Das Kriterium der *Nachvollziehbarkeit* soll sicherstellen, dass die Entscheidfindung des ADM-Systems offengelegt wird, so dass zumindest die betroffenen Personen in der Lage sind, die Würdigung der relevanten Fakten zu verstehen – und notfalls anzufechten.

Dem Rechtsrahmen liegt eine Einschätzung der Risiken zugrunde, welche von ADM-Systemen ausgehen. Betroffene Individuen, berechtigte NGOs sowie eine staatliche ADMS-Aufsichts sollen ein Recht auf Einsicht für Anwendungen enthalten, die ADM-Systeme einsetzen. Der Rechtsrahmen trägt auch dem Umstand Rechnung, dass das Risiko eines Systems sich mit der Zeit verändern kann. Zudem ist er kompatibel mit dem neuen Datenschutz-Konzept der Digitalen Gesellschaft.

Als Zivilgesellschaft haben wir heute nicht mehr die Möglichkeit zu entscheiden, ob wir den Einsatz von ADM-Systemen wünschen oder nicht. Sie sind bereits Realität. Wir haben aber die Wahl zu entscheiden, in welchen Bereichen wir uns durch ADM-Systeme unterstützen lassen wollen und in welchen Bereichen wir eine solche Unterstützung ablehnen.

2 Geltungsbereich

In den Anwendungsbereich unseres Vorschlags fallen ADM-Systeme, die Entscheidungen mit Hilfe von technischen Systemen vollständig automatisiert treffen oder zumindest unterstützen. Wir verzichten auf den umstrittenen Begriff der Künstlichen Intelligenz und übernehmen die folgende Definition von ADM-Systemen 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,[5] and the set of processes involved in implementing those tools.

Damit definieren wir ADM-Systeme nicht aus der Sicht der benutzten Technologie, sondern aus der Perspektive der Auswirkungen. Damit umgehen wir die Definitionsprobleme, mit denen beispielsweise die OECD zu kämpfen hatte (OECD 2024).

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 Regulierungsvorschlags 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. Die unter die Regulierung fallenden Entscheidungen müssen also als einzelne, diskrete Entscheidungen eine Auswirkung auf die Freiheiten, das Leben oder die Gesundheit, die wirtschaftliche oder gesellschaftliche Situati-

[4] Im Fokus unseres Regulierungsvorschlags stehen die Auswirkungen und Risiken der ADM-Systeme und nicht Verbote konkreter Technologien.

[5] 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.

on einzelner Personen oder Gruppen haben. Unter diese Definition soll auch Nudging durch ein ADM-System fallen, falls dessen Entscheide kumulativ eine gesellschaftliche Auswirkung aufweisen.

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.

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 Individuen und die Gesellschaft im Kontext der Schutzziele einschätzen und kategorisieren. Der Rechtsrahmen stellt dazu drei Kategorien zur Verfügung: «tiefes Risiko», «hohes Risiko» und «inakzeptables Risiko», welche von der Legislative definiert werden. 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 oder keinem Risiko» kein oder 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[6] 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 entgegnet diesem Defizit, indem er diese gesellschaftlichen Risiken anerkennt, und schlägt Abhilfe mittels kollektiver Rechtsbehelfe wie Sammelklagen und einem Verbandsklagerecht vor.

Der Rechtsrahmen unterscheidet zwischen ADM-Systemen, welche in der Privatwirtschaft eingesetzt werden, und solchen, die in Erfüllung öffentlicher Aufgaben verwendet werden (siehe Abschnitte 7.1 und 7.2). Für beide fordern wir ein Beschwerde-, resp. Klagerecht für betroffene Individuen, die staatliche ADMS-Aufsicht und berechtigte NGOs, um die korrekte Risiko-Klassifizierung gemäss Kapitel 6 und die Durchsetzung der damit verbundenen Pflichten zu garantieren.

Die neu zu schaffende ADMS-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, technischer und juristischer Expertise, soll sie finanziell und personell unabhängig und frei von Weisungen agieren.

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, die konkrete Umsetzung zur Einhaltung der Regeln innerhalb der durch den Rechtsrahmen definierten Parameter selbst zu gestalten. Doch dieser Freiheit sollen zusätzliche Pflichten entgegenhalten. Dazu zählen die Transparenzanforderungen, die Sorgfaltspflichten und die Beweislastumkehr sowie wirkungsvolle Sanktionen bei Missachtung.

Die genaue Funktionsweise eines ADM-Systems unterliegt meist dem Geschäftsgeheimnis.

[6] https://booksummaryclub.com/weapons-of-math-destruction-book-summary/

Entsprechend schwierig ist es für Aussenstehende, Beweise für das Risiko eines Systems zu beschaffen. Daher soll bei berechtigter Anschuldigung, das heisst, wenn ein Gericht auf die Klage eintritt, eine Beweislastumkehr gelten.[7] Dann muss das beschuldigte Unternehmen als erste Instanz in der Regresskette die korrekte Klassifizierung nachweisen. 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!».[8]

Die staatliche ADMS-Aufsicht unterstützt den Verantwortlichen[9] 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 Mitarbeitenden 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.

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.

4 Die gesellschaftliche Relevanz

Eine zentrale Einsicht ist, dass **automatisierte Entscheidungssysteme weder objektiv noch neutral sind**, denn sie repräsentieren stets die Werte ihrer Entwickler:innen sowie der Gesellschaft und können deshalb als sozio-ökonomischer Spiegel einer bestimmten Gesellschaft betrachtet werden. Problematisch wird dies, weil sich kulturelle und gesellschaftliche Werte rund um den Globus unterscheiden, wohingegen Technologien wie ADM-Systeme grenzüberschreitend oder global Einsatz finden können. Zudem können sich die Werte über die Zeit verändern, die einst definierten und langfristig eingesetzten Systeme müssen dies aber nicht zwingend umsetzen. Weiter sind ADM-Systeme durch ihre Funktion gebunden. Von Entwickler:innen bewusst oder unbewusst getroffene Design-Entscheide haben einen Effekt auf die Wirkungsweise des Systems und können sich negativ auswirken. Die Funktion beim Einsatz eines Systems bestimmt einen engen Handlungsrahmen, der oftmals nicht hinterfragt wird.

Ein Beispiel hier wäre der Einsatz von ADM-Systemen zur Reduktion von personellem Aufwand bei Massengeschäften in der Sozialhilfe. Diese Systeme fällen harte Entscheide über die verfügbaren Mittel der auf Sozialhilfe angewiesenen Menschen. Einerseits sind die Entscheidungsgrundlagen dieser Systeme fragwürdig, denn die von Entwickler:innen angestrebten Ziele sowie zugrundeliegenden gesellschaftliche Werte können sich im Zeitverlauf ändern. Zum anderen ist der Einsatz solcher Systeme grundsätzlich fragwürdig: Neuere, wissenschaftliche Studien liefern Hinweise, dass eher eine Erhöhung des personellen Aufwands die Sozialkosten insgesamt reduziert[10] (Eser Davolio 2020), was gegen eine Automatisierung der entsprechenden Verwaltung sprechen könnte.

Dieser Einschränkungen ist man sich aber oftmals nicht bewusst. Stattdessen betrachtet man die Resultate von ADM-Systemen allzu oft als objektiv richtig und korrekt. Im Kontext von Entscheidungspro-

[7] Die EU hat kürzlich die Beweislastumkehr bei Softwareprodukten eingeführt. Für technisch komplexe Produkte können Konsumentinnen und Konsumenten von einer Firma fordern, dass diese die ”notwendigen und verhältnismässigen”Beweise offenlegen muss.

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

[9] Bei diesem Begriff wird ausschliesslich die männliche Form verwendet, da wir uns an die Definition von Art. 5 lit. j DSG anlehnen.

[10] Siehe dazu https://www.zhaw.ch/de/forschung/forschungsdatenbank/projektdetail/projektid/1668/

blemen ist diese Sicht aber trügerisch, denn **für viele Entscheidungsprobleme existiert keine optimale Lösung**. Durch das Delegieren alltäglicher Aufgaben an ein ADM-System werden nun aber die Interaktionen mit diesem sowie dessen Resultate Teil der sozialen Realität. Die Soziologin Michele Willson beschreibt dies so: «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). Es entstehen Rückkopplungseffekte, die dazu führen, dass automatisierte Entscheidungssysteme sowie ihre (teils fehlerhaften oder ungenauen) Datengrundlagen sich stets verändern und Teil des sozialen Gefüges werden. Solche Effekte können sowohl beabsichtigt wie auch unbeabsichtigt sein.

Ein besonders problematischer Effekt stellt dabei Diskriminierung dar. Auf Datensätzen trainierte ADM-Systeme reproduzieren die darin implizit festgeschriebenen Diskriminierungspraktiken, etwa gegenüber sozial schwächeren Gruppen oder Menschen mit Behinderung. So würden auf historischen Daten trainierte Recruitment-Systeme Frauen oder Menschen mit einer Behinderung 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 durch den vermehrten, auch grenzüberschreitenden Einsatz verstärken.

Ein weiterer problematischer Effekt ist, dass automatisierte Entscheidungssysteme zunehmend den Informationsüberfluss kuratieren, zum Beispiel in Form von sogenannten «Vorschlagssystemen» oder als «Fakten-» und «Urheberrechtsprüfer». So werden Systembetreiber:innen ermächtigt, politische Botschaften und Positionen durch gezieltes Agenda-Setting selektiv zu verstärken. Dabei müssen diese Systeme (beispielsweise als Faktenprüfer) nicht zwingend auf Personendaten operieren. Vielen dieser Effekten von automatisierten Entscheidungssystemen ist gemein, dass sie oft im Verborgenen geschehen und ihre Auswirkungen erst spät oder durch weitere, indirekte Effekte bemerkt werden. Einsatz und Funktionsweise der Systeme sind oftmals nicht bekannt, da ihre Entwickler:innen sowie Betreiber:innen kein Interesse an einer Offenlegung haben.

Schliesslich lässt das Zusammenspiel unterschiedlicher ADM-Systeme 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. Diese **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 Menschen durch den Einsatz dieses 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 Regulierungsvorschlag für ADM-Systeme

Wir werden als Gesellschaft zunehmend mit den spezifischen Auswirkungen von ADM-Systemen konfrontiert. Daher fordern wir **eine substanzielle Erweiterung der bestehenden Gesetze** zur Berücksichtigung der Herausforderungen automatisierter Entscheidungssysteme oder sogar ein spezifisches **«ADMS-Gesetz»**, falls sich dies als vorteilhaft

erweist. Wir fordern **Transparenz** beim Einsatz von ADM-Systemen, um die bereits bestehenden Gesetze anwenden zu können, und **wirksame Strafen** bei Missachtung. Wir fordern **Erklärbarkeit** von ADM-Systemen, welche sich dem Menschen rasch und mit angemessenem kognitivem 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 menschliches Zutun – 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 Masse Kontrolle darüber auszuüben.

Wir fordern **klare Schutzziele, nämlich die Einhaltung der Grund- und Menschenrechte, die Wahrung der psychischen und physischen Gesundheit und Sicherheit der einzelnen Person, die Wahrung der Lebens- und Entwicklungschancen, sowie den Schutz der demokratischen Rechte und Prozesse**. Weiter müssen betroffene Personen und die Öffentlichkeit die Möglichkeit haben, die Einhaltung dieser Schutzziele effektiv zu kontrollieren und, wenn nötig, zu beanstanden und niederschwellig einzufordern. 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.

Die ADMS-Regulierung soll weder die Innovation hemmen noch die Unternehmen und die später im Detail beschriebene ADMS-Aufsicht bürokratisch unverhältnismässig belasten. Wir sprechen uns für einen breiten Rechtsrahmen aus, welcher die benötigte Regulierung generell einführt, aber es den einzelnen Wirtschaftssektoren ermöglicht, die effektivsten Methoden zur Umsetzung der Schutzziele selbst zu bestimmen. Unsere später detailliert 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.

Zur Unterstützung der verantwortlichen Entwicklung von ADM-Systemen sollen staatliche Massnahmen zur Förderung von Open-Source Libraries und Frameworks für ADMS- und KI-Entwickler:innen ergriffen werden.

Grundsätzlich können auch bestehende Gesetze mit einigen Änderungen ebenfalls auf ADM-Systemen Anwendung finden. So kann beispielsweise die Dispersion und Verwendung von Personendaten durch das Datenschutzgesetz geregelt werden.[11] Diskriminierungsverbote sind das Negativ zu der aufkommenden Fairness-Diskussion,[12] jedoch mit einem grossen Graubereich dazwischen[13], in dem sich der Hauptteil der realen Anwendungen wiederfinden wird. Das Arbeits- sowie das Datenschutzrecht verbietet einige Überwachungspraktiken, algorithmische und nicht algorithmische, am Arbeitsplatz.[14] Formaljuristisch könnte sich ein eigenes ADMS-Gesetz aus zwei Gründen als vorteilhaft erweisen: Erstens, weil die nötige Änderung in anderen Gesetzestexten eine gemeinsame Begriffserklärung sowie Definition

[11] Wir fordern eine Anpassung von Art. 21 Abs. 1 DSG (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).»

[12] Bei Fairness in 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 et cetera sind. Die genaue (mathematische) Formulierung von Fairness ist jedoch eine noch offene Debatte, und einige Definitionen widersprechen sich sogar.

[13] 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.

[14] Die Überwachung von Arbeitnehmer:innen ist in begrenztem Mass erlaubt, wie beispielsweise die Erfassung und Einhaltung der Arbeitszeit (Art. 46 ArG), Daten im Zusammenhang mit Eignung für das Arbeitsverhältnis et cetera. 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 beispielsweise 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).

der Kategorisierung und der Risiken benötigen und zweitens, weil die im Anschluss ausgeführte ADMS-Aufsicht schlecht anderswo definiert werden kann.

Juristische Personen können - gleich natürlichen Personen - von ADM-Systemen profitieren. Der Einsatz von KI beispielsweise in der Produktion, der Dienstleistungserbringung und so weiter ist mannigfaltig. Der Digitalen Gesellschaft sind bis zum Publikationszeitpunkt keine konkreten Beispiele bekannt, die auf eine Benachteiligung oder Diskriminierung von juristischen Personen hinweisen. Ungeachtet dessen ist sich die Digitale Gesellschaft bewusst, dass auch juristische Personen Nachteile erfahren können. Ob und in welchem Umfang dies geschehen kann und inwiefern die bisherigen Gesetze (wie beispielsweise das Wettbewerbsrecht) juristische Personen ungenügend schützen, kann zum jetzigen Zeitpunkt nicht abgeschätzt werden. Die Digitale Gesellschaft behält sich eine Stellungnahme zu diesem Thema zu einem späteren Zeitpunkt vor.

5.1 Die ADMS-Aufsicht

Die **staatliche ADMS-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 der Regulierung und die Kategorisierung von staatlichen und privatwirtschaftlichen ADM-Systemen.

Die ADMS-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 Einhaltung der ADMS-Regulierung und über die 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).

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 der EDÖB, 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 potenziell manipuliert werden. Massnahmen zur Unterbindung derartiger Angriffe gehören unserer Ansicht nach jedoch nicht in eine ADMS-Regulierung, sondern in ein generelles «**IT-Sicherheits-Gesetz**».

Das ADMS-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[15] in Betracht ziehen («reasonably foreseeable misuse», EU AI Act Art 9.2(b)). 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 Softwarefirmen 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.

[15] 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 (Adversarial Examples) et cetera.

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. Wir folgen dabei den Empfehlungen des «Gutachtens der Datenethikkommission» der Deutschen Bundesregierung (Seite 43ff)[16] sowie der ausführlichen Analyse der Grundrechtsimplikationen von Gesichtserkennungstechnologie in FRA 2019). 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. Potenzielle Pflichtverletzungen oder Fehlkategorisierungen werden a posteriori durch Strafen im Rahmen von Beschwerden und Klagen geahndet. Dieser Ansatz gibt den Betreiber:innen von ADM-Systemen die Möglichkeit, selbstverantwortlich, aber in einem klaren Rahmen, ADM-Systeme zu entwickeln und einzuführen. Die Selbstverantwortung wird durch die Strafmechanismen eingefordert und gestärkt.

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 im Kontext der Schutzziele aufgrund des Schadenspotentials sowie der Eintrittswahrscheinlichkeit für die Gesellschaft als Ganzes eruiert, während im individuellen Einzelfall das Risiko für die einzelnen Betroffenen betrachtet wird. Der Entscheidungsbaum für die Kategorien wird vom Gesetzgeber festgelegt, nicht von anderen Akteur:innen.

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 Schutzziele und Risiken für Individuen und die Gesellschaft als Ganzes

Beim Einsatz von ADM-Systemen ist der Schutz der Individuen und der Gesellschaft sicherzustellen. Die grundsätzlichen Schutzziele in Bezug auf Individuen sind die Einhaltung der Grund- und Menschenrechte. Wir orientieren uns hierbei an den Schutzzielen unseres Datenschutz-Konzepts (s. Digitale Gesellschaft 2023), welche primär auf den Schutz von Individuen ausgelegt sind, jedoch auch gesellschaftliche Risiken beinhalten, die nicht direkt auf der Kumulierung von Individualrisiken basieren:

- Schutz vor Manipulation
- Schutz vor Diskriminierung
- Schutz vor Überwachung und Recht auf Anonymität
- Schutz vor Beeinträchtigung der Gesundheit sowie der Lebens- und Entwicklungschancen
- Recht auf Transparenz und Pflicht zur Sorgfalt
- Recht auf Vergessen
- Schutz der offenen Gesellschaft und freien Demokratie

Unter Manipulation zu verstehen ist die absichtliche, gezielte und in der Regel verdeckte Einflussnahme auf die Entscheidung einer anderen Person, um deren Selbstkontrolle und Entscheidungskraft zu unterlaufen. ADM-Systeme erlauben dabei eine hoch automatisierte und individuelle Ansprache von Personen. Die Manipulation kann zu einem Nachteil für die betroffene Person führen. Sie zielt unter Ausnützung menschlicher Schwächen auf eine Steuerung des Verhaltens von Individuen oder Gruppen. Vulnerable Menschen sind besonders stark gefährdet.

Diskriminierung findet dann statt, wenn ADM-Systeme mit ihren Entscheidungen Menschen oder Gruppen aufgrund von Eigenschaften wie ethnischer Zugehörigkeit, Hautfarbe, Geschlecht, Klassenzugehörigkeit, sexueller Orientierung et cetera benachteiligen. In der Regel ist der zugrundeliegende «Bias» bereits in den Trainingsdaten vorhanden und wird durch die Automatisierung von Entscheidungen fortgeschrieben oder sogar verstärkt. Beispiele für Diskriminierung werden im Dossier der Fachgruppe Tracking & Profiling der Digitalen Gesellschaft gegeben.

[16] vgl. Datenethikkommission der Bundesregierung 2019

Wird der Schutz vor Überwachung und das Recht auf Anonymität verletzt, werden Menschen daran gehindert, ihre Identität zu entwickeln. Es kann ausserdem zu «Chilling Effects» kommen, das heisst allein durch die Möglichkeit der Überwachung sehen Menschen davon ab, an Demonstrationen und Kundgebungen teilzunehmen, mit äusserst negativen Auswirkungen auf den demokratischen Willensbildungsprozess einer Gesellschaft (dies ist der Hauptgrund, warum netzpolitische NGOs mit Nachdruck gegen biometrische Identifikation und Gesichtserkennung auf öffentlich zugänglichem Grund kämpfen).

ADM-Systeme, die Entscheidungen in den Bereichen Soziales, Rechtsprechung und Strafverfolgung, Bildung sowie im täglichen wirtschaftlichen Leben treffen oder unterstützen, können massive Auswirkungen auf die Entwicklungschancen von Individuen haben (beispielsweise ein nicht gewährter Studienplatz oder Kredit). Für Beispiele und weitere Ausführungen sei hierzu wiederum auf das Dossier der Fachgruppe Tracking & Profiling der Digitalen Gesellschaft verwiesen.

Das Recht auf Transparenz kann auf mehreren Ebenen verletzt werden. Zum einen dadurch, dass Personen nicht bewusst ist, dass sie betreffende Entscheidungen automatisiert getroffen werden. So müssen gemäss dem neuen Datenschutzgesetz beispielsweise nur vollständig automatisierte Entscheidungen neu ausgewiesen werden. Zum anderen ist es für Individuen generell intransparent, welche Daten verwendet werden oder mit welchen Modellen und Koeffizienten (siehe Glossar) jeweils Entscheide getroffen werden. Auch hierzu verweisen wir auf das Dossier der Fachgruppe Tracking & Profiling der Digitalen Gesellschaft.

Das Recht auf Vergessen wird in der Regel selten im Zusammenhang mit ADM-Systemen diskutiert, ist aber auch dort relevant. Wie lange dürfen Daten der Vergangenheit für automatisierte Entscheidungen berücksichtigt werden, wie beispielsweise bei der Berechnung von Credit-Scores. Dies ist insbesondere dann herausfordernd, falls ein ADM-System auf diesen Daten bereits trainiert wurde und korrigiert werden müsste.

Die offene Gesellschaft und freie Demokratie sind im Kontext von ADM-Systemen da in Gefahr, wo Menschen oder ganze Gruppen aufgrund bestimmter Eigenschaften diskriminiert werden (siehe Schutzziel Diskriminierung) und wo demokratische Prozesse gestört werden (siehe Schutzziel Schutz vor Überwachung und Recht auf Anonymität). Die Demokratie ist ausserdem dann in Gefahr, wenn Menschen oder ganze Gruppen manipuliert werden - beispielsweise in sozialen Medien. Oder wenn Botschaften zielgerichtet und individuell auf einzelne Personengruppen zugeschnitten werden, und keine Transparenz darüber besteht, welche Informationen ausgespielt werden. Dadurch wird der Diskursraum fragmentiert und damit eine inhaltliche Debatte erschwert.

6.2 Beurteilungskriterien

Risiken werden als Kombination der Schwere des möglichen Schadens und der Wahrscheinlichkeit, dass der Schaden eintritt, verstanden. Verkürzt gesagt Risiko = Schadensausmass * Eintrittswahrscheinlichkeit des Schadens. Der Schaden wird mit Blick auf die Schutzziele (vgl. 6.1) bewertet. Die Wahrscheinlichkeit, dass ein Schaden eintritt, kann sich aus mehreren Faktoren ergeben. Eine Evaluationsmöglichkeit wäre beispielsweise, wie wahrscheinlich eine problematische Situation entsteht (exposure) und wie wahrscheinlich diese vor Schadenseintritt korrigiert werden kann. Unter Umständen ist auch die Rückgängigmachung des Schadens zu berücksichtigen, beispielsweise spätere Geldüberweisung nach anfänglicher Sperrung.

Für die Regulierung sollen ADM-Systeme entsprechend ihren Risiken kategorisiert werden. Für die Kategorisierung übernehmen und ergänzen wir einige der Konzepte aus dem AI Act der EU Kommission (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 nachteiligen Auswirkungen?
- Worin besteht das potenzielle Ausmass solcher Schäden, solcher Beeinträchtigungen oder solcher nachteiligen 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 der einsetzenden Entität 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 entstehen, und welche Massnahmen werden dagegen getroffen?

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 (das heisst 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 mit Blick auf die Schutzziele 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.

Um die Innovation nicht zu hemmen, ist es gleichzeitig auch wichtig, dass die Rechtssicherheit bei der Selbstdeklaration gewahrt ist. Die Verwaltungssanktionen sollen nicht Unternehmen treffen, die die Selbstdeklaration sorgfältig sowie nach bestem Wissen und Gewissen vornehmen. Diesem Umstand soll insbesondere durch Merkblätter der ADMS-Aufsicht Rechnung getragen werden, die die Kriterien zur Selbstdeklaration konkretisieren und Beispiele aufführen (vgl. etwa im Datenschutzrecht: EDÖB 2023).

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 («Tiefes Risiko») fallen Systeme,

- die ein tiefes Risiko für die Gesellschaft darstellen und
- die für Individuen
  - kein oder nur ein geringes Risiko für die Schutzziele 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.

Beispiele für ADM-Systeme in dieser Kategorie sind die automatische Überprüfung von Lebensmittelverpackungen auf Korrektheit direkt nach der Produktion oder ADM-Systeme zur Vorhersage von Pollenbelastung, welche eine grosse Hilfe für Menschen mit Allergien darstellen, aber im Falle von Fehlverhalten nur vernachlässigbare Auswirkungen zeigen.

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 Schutzziele darstellen.

Die Systeme in dieser Kategorie sind dadurch gekennzeichnet, dass ihrem (positiven) Nutzen ein signifikantes potenzielles negatives Potential gegenübersteht. Das Schadenspotential ist dabei noch akzeptabel (oder mitigierbar), ansonsten würde ein solches System in der nächsthöheren 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.

ADM-Systeme, 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 (beispielsweise 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, beispielsweise in der medizinischen Diagnostik. Diese würden damit als «inakzeptabel» kategorisiert. Solange diese Systeme als Unterstützung eingesetzt werden und die endgültige Entscheidung von einer Fachperson getroffen wird, ist eine Rückstufung zu «hohem Risiko» sinnvoll. Jedoch ist der Übergang von häufig eingesetzten Recommender-Systemen als Unterstützungssysteme bis zum unhinterfragten Akzeptieren dieser Vorschläge fliessend, was anfängliche Unterstützungssysteme zu De-facto-Entscheidern 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 Schutzziele oder
  - irreversible und schwerwiegende Auswirkungen darstellen.

In diese Kategorie gehören also Systeme, deren potenzieller Schaden so gross ist, dass er nicht riskiert werden darf. Bei vielen Systemen in dieser Kategorie ist der Schaden ausserdem bekannt und dokumentiert, und damit nicht mehr potenziell, sondern kann zuverlässig erwartet werden. Die Entscheidungen in dieser Kategorie sind ausserdem weder umgehbar (beispielsweise biometrische Massenüberwachung) noch revidierbar, sie sind zudem oft auch nicht überprüfbar (beispielsweise 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. Der Einsatz solcher Systeme wird verboten.

Beispiele für inakzeptable Auswirkungen für die Gesellschaft sind die bereits oben genannte biometrische Massenüberwachung (inkl. Gesichtserken-

nung[17]), 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 demokratisch 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.[18]

7 Sorgfalts- und Transparenzpflichten

Grundsätzlich ist die Entität, die aus Sicht der Betroffenen das ADM-System einsetzt, (zum Beispiel die betreibende Firma) 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:innen 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:innen von jeglicher Verantwortung mittels der AGB (Allgemeinen Geschäftsbedingungen) 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[19] wie bei medizinischen Produkten als sinnvoll, beispielsweise bei einem automatischen Defibrillator (AED). Sonst besteht die Gefahr, dass die Rechenschaftspflicht grossflächig an Zertifikatsaussteller ausgelagert wird.

Eine Möglichkeit, die Sorgfaltspflichten zu unterstützen, wären Folgeabschätzungen (Impact Assessments), welche die möglichen Folgen der Entwicklung und Einsatzes eines ADM-Systems beleuchten und die Basis für durchdachte Risiko-Mitigationsstrategien legen sowie als Indiz für den verantwortungsvollen Umgang mit der Technologie dienen können. Zum Verfassungszeitpunkt dieses Papiers sieht die Digitale Gesellschaft jedoch keinen direkten Vorteil darin, diese Instrumente verpflichtend vorzuschreiben.

Die Einordnung in eine Risikokategorie ist von der Betreiberin oder vom Betreiber zu dokumentieren. Für ADM-Systeme mit «tiefem Risiko» reicht eine formlose, aber nachvollziehbar begründete Einordnung aus. Für ADM-Systeme mit «hohem Risiko» ist eine systematische Analyse nötig, die beispielsweise im Rahmen eines Risikomanagementprozesses erfolgen kann, der für bestimmte Produktklassen wie beispielsweise Medizinprodukte ohnehin existieren muss.

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 ADMS-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

[17] beispielsweise https://gesichtserkennung-stoppen.ch, https://reclaimyourface.eu

[18] 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); siehe Crawford et al. zur Kritik an Techniken der automatisierten Emotionserkennung (vgl. Crawford 2021)

[19] Dies bedeutet eine klar überprüfbare Funktionsweise des selbständig entscheidenden Systems.

1. darauf hinweist, dass ein ADM-System eingesetzt wird,[20]
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 mehreren ADM-Systemen (unter Umständen von verschiedenen Herstellerfirmen) besser sichtbar wird. Des Weiteren fordern wir eine periodische und kontinuierliche Überprüfung des Risikos (das heisst der Einstufung der Kategorie) sowie der Dokumentation bezüglich der Transparenzpflichten, vor allem bei selbständig weiterlernenden Systemen.[21]

Bei der im vorherigen Absatz erwähnten **Datenqualität** geht es darum, dass die verwendeten Daten entweder mit der Realität übereinstimmen,[22] dass darauf basierende Systeme also möglichst fehlerlos funktionieren, oder dass sie nur in beabsichtigten und allgemein nützlichen Aspekten verändert wurden, beispielsweise um Diskriminierung vorzubeugen.

Bezüglich der **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.[23]

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 et cetera 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 die Schutzziele, insbesondere die Gesundheit, Sicherheit oder Grundrechte, des Einzelnen oder der Gesellschaft einschätzen zu können.

Eine formelle Regulierung 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).

7.2 In Erfüllung eines öffentlichen Auftrags

Für staatliche Akteure ergibt sich bereits aus bestehenden Regelungen ein höherer Standard für Nachvollziehbarkeit und Transparenz, beispielsweise aus Öffentlichkeitsrecht und Strafprozessordnung. Diese Standards gelten als Minimum auch beim Einsatz von ADM-Systemen. Neben denselben Transparenzpflichten wie für privatwirtschaftliche ADM-Systeme (Informationen über die Datenherkunft und -Qualität, Features und Zweck) fordern wir bei ADM-Systemen staatlicher Akteure die Offenlegung der Koeffizienten (siehe Glossar) in standardisiertem Format[24] wie folgt:

[20] mit Anpassung von Art. 21 Abs. 1 nDSG (streichen von «ausschliesslich»)

[21] 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.

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

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

[24] Dies macht die automatische Auswertung einfacher.

- Für ADM-Systeme, die auf Nicht-Personendaten basieren, sollen die Daten als OpenData 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 ADMS-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.

Die Leitlinien des Bundes für Künstliche Intelligenz und deren Monitoring zeigen die Relevanz des Themas für die Bundesverwaltung.

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 einen entsprecheden 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.[25]. Eine Verankerung in der ZPO wäre zwar zu begrüssen, aber wir fordern die Einführung von kollektiven Rechtsdurchsetzungsmitteln unabhängig vom Ausgang dieser Beratungen.

Die ADMS-Aufsicht soll Verstösse gegen die Regulierung 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 ADMS-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

[25] Über die Einführung von allgemeinen kollektiven Rechtsdurchsetzungsmitteln in der Zivilprozessordung (ZPO) wird zum Publikationszeitpunkt dieses Dokuments im Parlament debattiert

der Regel bei Verletzungen der oben aufgeführten Sorgfalts- und Transparenzpflichten um ein Organisationsverschulden. Die ADMS-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 ADMS-Aufsicht stehen insbesondere folgende Instrumente zur Verfügung: Sie sammelt Reklamationen, sie kann Einsicht verlangen und Sanktionen und Verfügungen erlassen. 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 ADMS-Aufsicht führen, falls die Verfügungen der ADMS-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 ADMS-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. Im Falle einer ernsthaften Anschuldigung, das heisst, falls ein Gericht die Beschwerde oder Klage als zulässig anerkennt, fordern wir daher die **Beweislastumkehr**, so dass beschuldigte Entitäten (Betreiber:innen der ADM-Systeme) hinreichend belegen müssen, dass sie die Kategorisierungsvorgaben, Sorgfalts- oder Transparenzpflichten nicht verletzt haben. Diese Beweislastumkehr ist einer der Gegen-Pflichten zum Vertrauensvorschuss der ADM-Selbstkategorisierung der Unternehmen.

Wie bei ähnlichen Technologie- und Softwareprodukten, soll es der Betreiberin möglich sein für die Schäden Schadenersatz zu fordern, die auf seine Zulieferer, wie beispielsweise Entwicklerinnen oder Systembetreiber, zurückzuführen sind. Für die Betroffenen bleibt die Betreiberin aber immer selbst verantwortlich.

8.2 In Erfüllung eines öffentlichen Auftrags

Grundsätzlich sollen die gleichen Kontrollen, Massnahmen und Sanktionen wie gegen private Entitäten möglich sein. Wie das Verhältnis zwischen der ADMS-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 ADMS-Aufsicht im kantonalen Verfahren immer Parteirechte zukommen.

Auch hier soll es, wie bei ähnlichen Technologie- und Softwareprodukten üblich, für die Behörde als Betreiberin und Auftraggeberin möglich sein für die Schäden Schadenersatz zu fordern, die auf seine Zulieferer, wie beispielsweise Entwicklerinnen oder Systembetreiber, zurückzuführen sind. Für die Betroffenen bleibt die Behörde aber immer direkt verantwortlich.

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 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. Potenzielle 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 diversen Diskurs-Kreisen besteht die Vision, dass sich alle sozialen 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 potenziellen **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 kontinuierlicher **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.

A Regulierungsvorschläge für ADM-Systeme, 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) und 2022 aktualisiert (vgl. ACM 2022). Die ACM ist der Berufsverband der US-amerikanischen Informatiker:innen und somit die Organisation, der viele der 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).

Viele 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 2024), der im Juni 2024 adoptiert 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).

Der ÄI Bill of Rights"(White House Office of Science and Technology Policy 2022) in den USA ist streng genommen kein eigener Regulierungsversuch, sondern formuliert und schärft Grundrechte im Kontext der Künstlichen Intelligenz, wie beispielsweise das Recht auf Schutz vor algorithmischer Diskriminierung, Recht auf Transparenz und Erklärung oder das Recht auf menschliche Intervention. Obwohl der Bill «AI» im Titel trägt, treffen viele seiner Aussagen besonders auf ADM-Systeme zu. Der Accountability Act ist ein weiterer Gesetzesvorschlag aus den USA. Der Geltungsbereich dieses Vorschlags sind kritische Entscheidungen über Konsument:innen, beispielsweise in den Bereichen Bildung, Arbeit, oder Gesundheitswesen. Der Vorschlag hat zum Ziel, negative Auswirkungen auf Konsument:innen zu minimieren und schlägt diverse Rechte für Konsument:innen vor, beispielsweise Kennzeichnungspflicht der ADM-Systeme, Opt-out-Möglichkeiten und Widerspruchs- und Korrekturmöglichkeiten.

Auch in der VR China gibt es Vorschläge zum Umgang mit KI und ADM-Systemen (vgl. National New Generation Artificial Intelligence Governance Specialist Committee). Dieser Vorschlag formuliert ethische Richtlinien im KI-Bereich, Die Richtlinien sind in Grundnormen wie Förderung des menschlichen Wohlergehens, Förderung von Fairness und

Gerechtigkeit, Schutz der Privatsphäre oder Stärkung der Verantwortlichkeit gruppiert.

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.

B Quellenverzeichnis

B.1 Quellen hinsichtlich Regulierungsvorschläge von ADM-Systemen im europäischen sowie interkontinentalen Kontext

- ACM (2017). Statement on Algorithmic Transparency and Accountability (Zugriff 13.0.7.2024). *ACM U.S. Public Policy Council.*
- ACM (2022). Statement on Principles for Responsible Algorithmic Systems (Zugriff 13.0.7.2024). *Europe/U.S. Public Policy Council.*
- BAKOM (2022). Monitoring der Leitlinien «Künstliche Intelligenz» für den Bund (Zugriff 13.0.7.2024). *Bundesamt für Kommunikation BAKOM*.
- Bundesrat (2020). Die Leitlinien des Bundes für Künstliche Intelligenz (Zugriff 13.0.7.2024). *Der Bundesrat*.
- CAHAI (2021). A legal framework for AI systems (Zugriff 13.0.7.2024). *Ad hoc Committee on Artificial Intelligence of the Council of Europe*.
- Datenethikkommission der Bundesregierung (2019). Gutachten der Datenethikkommission der Bundesregierung (Zugriff am 16.12.2019). *Berlin: Bundesministerium des Innern, für Bau und Heimat.*
- EU AI Act (2024). REGULATION (EU) 2024/… OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). OJ 2024 L .... (Zugriff am 28.06.2024). *The european parliament*.
- 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. *Das europäische Parlament*.
- European Commission adoption consultation (2021). Artificial Intelligence Act (Zugriff 13.0.7.2024). *EDRi*.
- European Union Agency for Fundamental Rights (2019). Facial recognition technology: fundamental rights considerations in the context of law enforcement (Zugriff am 08.02.2022). *European Union Agency for Fundamental Rights*.
- National New Generation Artificial Intelligence Governance Specialist Committee (2021). Ethical Norms for New Generation Artificial Intelligence Released (Zugriff 13.0.7.2024). *Center for security and emergin technology*.
- Repräsentantenhaus (2022). H.R.6580 - Algorithmic Accountability Act of 2022 (Zugriff 13.0.7.2024). *117th Congress (2021-2022) of the United States of America*.
- Richardson, Rashida et al. (2019). Confronting Black Boxes: A Shadow Report of the New York City Automated Decision System Task Force (Zugriff 13.0.7.2024). *AI Now Institute*.
- Thouvenin, Florent et al (2021). Ein Rechtsrahmen für Künstliche Intelligenz. *Digital Society Initiative Universität Zürich.*
- Verbraucherzentrale Bundesverband (2021). Artificial Intelligence needs Real-World Regulation (Zugriff 13.0.7.2024). *Verbraucherzentrale Bundesverband*.
- White House Office of Science and Technology Policy (2022). Blueprint For An AI Bill Of Rights Making Automated Systems Work For The American People. *The White House*.

B.2 Weitere Quellen

- AI Now Insititute (2018). Automated Decision Systems - Examples of Government Use Cases (Zugriff am 08.02.2022). *AI Now Insititute*.
- Assion, Simon (2014). Überwachung und Chilling Effect. *«Überwachung und Recht», Tagungsband zur Telemediucs Sommerkonferenz 2014, epubli GmbH, Berlin.*
- Bürgi, Urs (2022). Arbeitnehmerüberwachung (Zugriff am 25.01.2022). *Law Media*. Urheber: Bürgi Nägeli Rechtsanwälte.

- Crawford, Kate (2021). Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence. *Yale University Press, New Haven*
- Digitale Gesellschaft (2023). Datenschutz-Konzept der Digitalen Gesellschaft (Zugriff 24.06.2024). *Digitale Gesellschaft*. Vorgestellt am Datenschutzfestival vom 03.11.2023.
- Eidgenössischer Datenschutz- und Öffentlichkeitsbeauftragter (EDÖB) (2023). Merkblatt zur Datenschutz-Folgenabschätzung (DSFA) nach den Art. 22 und 23 DSG (Zugriff am 25.20.2023). *Eidgenössischer Datenschutz- und Öffentlichkeitsbeauftragter (EDÖB).*
- Ensign, Danielle, et al. (2018). Runaway Feedback Loops in Predictive Policing. *FAT 2018*.
- Eser Davolio, M. et al. (2020). Auswirkungen der Falllastreduktion in der Sozialhilfe auf die Ablösequote und Fallkosten: Entschleunigung zahlt sich aus (Zugrifff 13.07.2024). *Schweizerische Zeitschrift für Soziale Arbeit*, 2019(25).
- Fanta, Alexander (2020). Datenschutzbehörde stoppt Jobcenter-Algorithmus (Zugriff am 7.8.2022). Netzpolitik.org, 21.8.2020.
- FDA (2023). Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices | FDA (Zugriss am 28.10.2023). fda.gov.
- 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. *NZZ am Sonntag*, 19.09.2021.
- NiederlandeNet (2020). RECHT: Gericht verbietet Betrugsbekämpfung mit Hilfe des Computerprogramms SyRi wegen Eingriffs in Privatsphäre (Zugriff am 7.8.2022). *NiederlandeNet*, WWU Münster, 5.2.2020.
- OECD (2024). Explanatory memorandum on the updated OECD definition of an AI system. *OECD Artificial Intelligence Papers*, No. 8.
- Orwat, C. (2019). Diskriminierungsrisiken durch Verwendung von Algorithmen. I*nstitut für Technikfolgenabschätzung und Systemanalyse (ITAS)*, Karlsruher Institut für Technologie (KIT).
- O'Neil, Cathy (2016). Weapons of Math Destruction. *New York: Crown.*
- Penney, Jonathon (2016). Chilling Effects: Online Surveillance and Wikipedia Use (Zugrifff 13.07.2024). *Berkeley Technology Law Journal*, Vol. 31, No. 1, p. 117, 2016.
- Public Code, Public Money (n.d.). 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 (Zugriff am 08.02.2022), *blogs.microsoft.com.*
- Tufekci, Zeynep (2018). YouTube, the Great Radicalizer (Zugriff am 8.2.2022), *The New York Times*, 10.3.2018.
- Willson, Michele (2017). Algorithms (and the) everyday. *Information, Communication & Society*, 20:1, 137-150, DOI: 10.1080/1369118X.2016.1200645.

B.3 Bildnachweis

- Bild auf Titelseite: Foto - Robynne Hu, Unsplash- Lizenz

C 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 et cetera 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.
- **Foundation Model:** Dies sind KI- und ADM-Systeme, die durch die Generalität ihrer Bearbeitungs- und Ausgabemöglichkeiten für eine Vielzahl an bekannten und auch derzeit noch unbekannten Aufgaben eingesetzt werden können.
- **Feedback-Loop**: Feedback-Loops (Rückkopplungsschleifen) beschreiben im ADMS-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 verrechnet 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).
- **Künstliche Intelligenz, KI:** Dies sind Systeme oder Algorithmen, welche (komplexe) Aufgaben von Menschen übernehmen können. Aufgrund der Umstrittenheit und Breite des Begriffs verzichten wir hier auf eine Definition und referenzieren im Kontext dieses Dokuments auf ADM-Systeme.
- **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 trainierten ADM-Systems zu evaluieren.

D Änderungstabelle

Tabelle mit den Änderungen von Version 1.0 zu Version 2.0 dieses Dokuments.

Kapitel (Version 1.0)	Änderungen
Generell	Versionsnummer auf 2.0 erhöht, zusätzliche Autoren hinzugefügt, ehemalige Autoren entsprechend erwähnt.
0. Executive summary	Neu hinzugefügt.
1. Einleitung	Neuschreibung der Einleitung.
2. Geltungsbereich	Erwähnt, warum wir den Begriff KI nicht benutzen. Hinweis auf Nudging ergänzt.
3. Zusammenfassung Rechtsrahmen	Beweislastumkehr genauer erläutert. Gegengengewicht Freiheit bei der Selbstklassifizierung vs. erweiterte Pflichten hervorgehoben.
4. Die gesellschaftliche Relevanz	Umgeschrieben, Beispiel ADM-Systeme bei Sozialbehörden angefügt, Fall Winterthur erwähnt und zitiert.
5. Ein Regulierungsvorschlag für automatisierte Entscheidungssysteme	Titel geändert. Ausführung zu juristischen Personen hinzugefügt. Ausführung zur staatlichen Förderung von open-source Libraries und Tools eingefügt.
6. Kategorisierung	Kapitel 6.1 umgeschrieben und an den Schutzziele ausgerichtet, vorherige Versionen entfernt. Definition von ”Risikoïn 6.2 integriert. In 6.2 ”Richtlinien aus dem AI Act der EU Kommission”durch ”Konzepte aus dem AI Act der EU Kommission ërsetzt. In Unterkapitel 6.3 wurde ein kurzer Abschnitt zur Gewährleistung von Rechtssicherheit für Unternehmen (insb. durch Merkblätter der ADMS-Aufsicht) hinzugefügt.
7. Sorgfalts- und Transparenzpflichten	Einleitungsabsatz klarer formuliert und Betreiber als Beispiel genannt. Folgeabschätzungen (Impact Assessment) diskutiert. Hinweis auf Dokumentation der Einordnung in eine Risikokategorie und auf Risikomanagement ergänzt. Hinweise auf bestehende Regelungen für staatliche Akteure ergänzt. Hinweis auf Leitlinien für KI und Evaluation ergänzt.
8. Kontrolle, Massnahmen und Sanktionen	Erwähnung des aktuellen Standes der Revision der ZPO in einer Fussnote. Streichung des Vergleichs der ADMS-Aufsicht mit der FINMA. Beweislastumkehr genauer erläutert. Regresskette in 8.1 und 8.2 erwähnt. Verantwortung des Betreibers bleibt.
9. Einige Anregungen für die Zukunft	Kleinere Änderungen.
A Feedback-Loops	Gelöscht.
B Regulierungsvorschläge für ADMS, Künstliche Intelligenz und Algorithmen	Anhang B aktualisiert, wurde zu Anhang A umbenannt. Titel leicht angepasst.