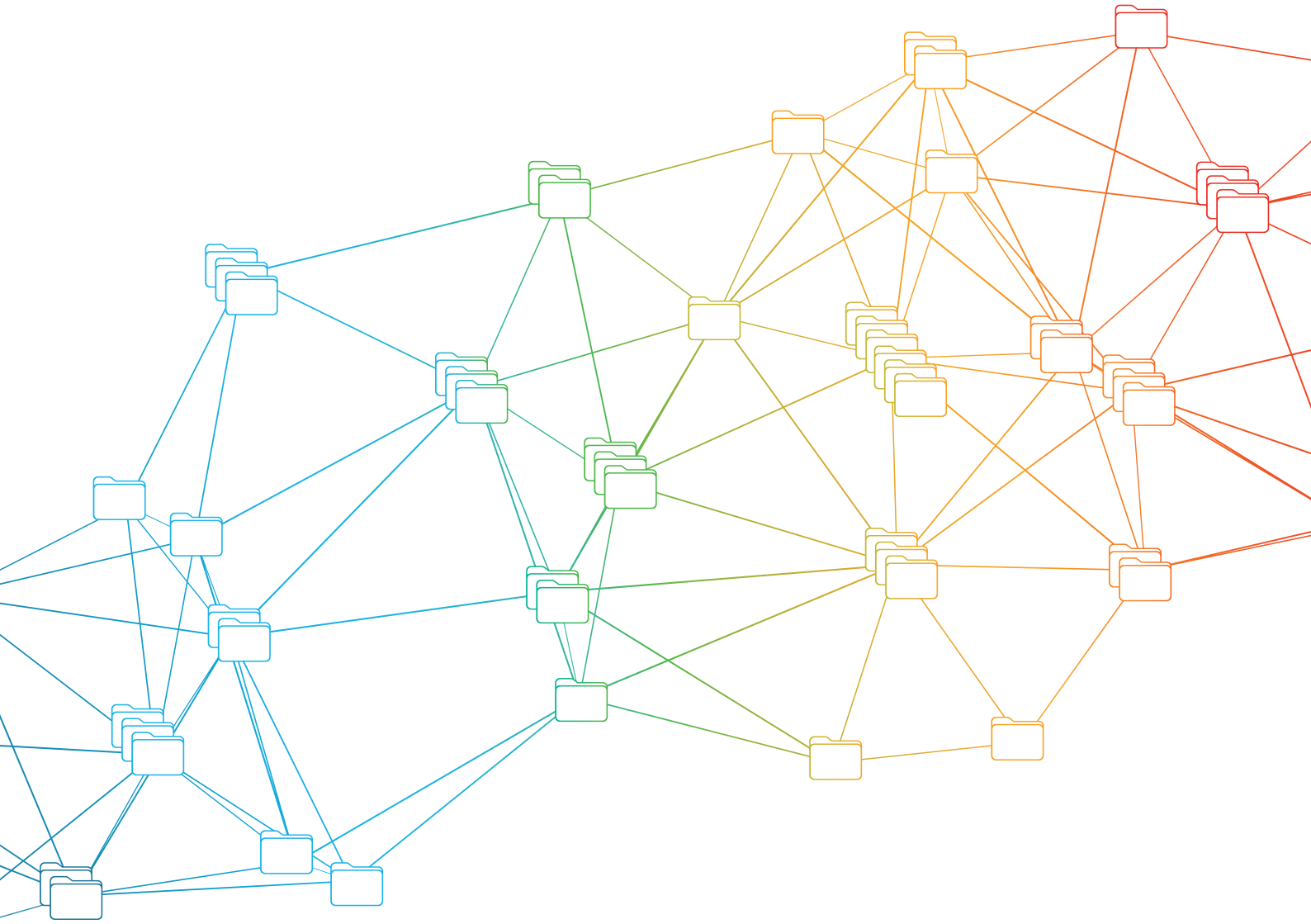
# Automatisierte Entscheidungssysteme im öffentlichen Sektor

## Ein Impact-Assessment-Tool für die öffentliche Verwaltung

Von Michele Loi, in Zusammenarbeit mit Anna Mätzener, Angela Müller und Matthias Spielkamp



ALGORITHM WATCH

ALGORITHM WATCH / CH

# Zusammenfassung und politische Handlungsempfehlungen

Beim Einsatz von automatisierten Entscheidungsfindungssystemen (ADMS) im öffentlichen Sektor muss es in erster Linie darum gehen, dass diese den Menschen und der Gesellschaft nützen und nicht schaden, dass sie die Autonomie der Betroffenen fördern und Grundsätze der Gerechtigkeit und Fairness achten. Angesichts des einzigartigen Kontextes, in dem Behörden handeln, und der besonderen Verantwortung, die sie wahrnehmen, muss ihr Einsatz von ADMS von einer **systematischen Bewertung potenzieller Auswirkungen** begleitet werden, die **Transparenz und Rechenschaftspflicht gegenüber den Betroffenen** gewährleistet.

In Zusammenarbeit mit der Universität Basel hat AlgorithmWatch im Auftrag des Kantons Zürich eine Studie über den Einsatz von Künstlicher Intelligenz in der öffentlichen Verwaltung durchgeführt. Als Ergebnis dieser Studie haben wir ein konkretes und praktikables Instrument zur Folgenabschätzung entwickelt, das für die Bewertung spezifischer ADMS durch öffentliche Behörden auf verschiedenen Ebenen eingesetzt werden kann.

Im vorliegenden Papier skizzieren wir zunächst die ethischen Grundlagen unseres Ansatzes, aus denen konkrete Fragen für die Bewertung von ADMS abgeleitet werden. Wenn das Ziel darin besteht, die Grundsätze der Schadensvermeidung, der Autonomie, der Gerechtigkeit und Fairness sowie der Benefizienz zu wahren, führt kein Weg daran vorbei, Transparenz, Kontrolle und Rechenschaftspflicht zu gewährleisten. Letztere sind keine Ziele an sich, sondern die notwendigen Mittel, um die eine verantwortungsvolle Nutzung von ADMS zu gewährleisten.

Diese Grundlage operationalisieren wir in einem zweistufigen Verfahren zur Folgenabschätzung. Es ermöglicht eine Triage von ADMS anhand von Risikosignalen und zeigt an, ob ein System zusätzlichen Transparenzanforderungen unterworfen werden muss. Wenn dies der Fall ist, müssen die Behörden einen umfassenden Transparenzbericht vorlegen. Abschließend veranschaulichen wir die Anwendung des Folgenabschätzungsinstruments anhand eines fiktiven Beispiels.

## Politische Handlungsempfehlungen

— Wenn ADMS im öffentlichen Sektor eingesetzt werden, sollten die Behörden verpflichtet werden, die potenziellen Risiken ihres Einsatzes systematisch zu bewerten und transparent zu machen. Diese Risiken können nicht pauschal, sondern nur durch eine Einzelfallanalyse ermittelt werden. Daher sollten die Behörden verpflichtet sein, sowohl vor und während der Einführung von ADMS eine Folgenabschätzung durchzuführen, als auch während des Einsatzes.

— Das hier entwickelte zweistufige Folgenabschätzungsverfahren ist ein praktikables und leicht einsatzbereites Instrument, um Transparenz über potenzielle Risiken zu schaffen, das auf sieben ethischen Grundsätzen beruht. Es ermöglicht eine Triage von ADMS und gibt an, ob ein bestimmtes System Risikosignale aufweist und damit zusätzlichen Transparenzanforderungen unterworfen werden muss. Ist dies der Fall, müssen die Behörden dafür sorgen, dass ein Transparenzbericht erstellt wird, der eine

Bewertung des Systems und seines Einsatzes während seines gesamten Lebenszyklus' ermöglicht. Transparenz allein gewährleistet noch nicht die Übereinstimmung mit den ethischen Anforderungen, ist aber eine notwendige Voraussetzung dafür.

— Die öffentliche Verwaltung sollte ein öffentliches Register für alle ADMS führen, die im öffentlichen Sektor eingesetzt werden. Neben Informationen über den Zweck, das zugrundeliegende Modell und die Entwickler*innen und Anwender*innen des Systems sollte dieses Register auch die Ergebnisse der Folgenabschätzung, d.h. gegebenenfalls den Transparenzbericht, enthalten. In Fällen, in denen es berechtigte Gründe gibt, den Zugang zum Transparenzbericht zu beschränken, sollten die Stellen aufgeführt werden, denen gegenüber er offengelegt wird.

# Inhalt

**A. Einleitung** ..... 4

**B. Einführung** ..... 7

**I. Ethische Richtlinien für den öffentlichen Sektor** ..... 7

**II. Zweistufiges Beurteilungsverfahren** ..... 9

**C. Sieben Grundsätze** ..... 12

**I. Ethische Grundsätze** ..... 12

1. Schadensvermeidung ..... 12
2. Gerechtigkeit und Fairness ..... 12
3. Autonomie ..... 14
4. Benefizienz ..... 16

**II. Instrumentelle und aufsichtsrechtliche Grundsätze** ..... 16

1. Kontrolle ..... 16
2. Transparenz ..... 19
3. Rechenschaftspflicht ..... 22

**D. Checklisten** ..... 24

**I. Einleitung** ..... 24

**II. Checkliste 1: Triage-Checkliste für KI-Systeme** ..... 25

1. Einleitende Bemerkungen ..... 25
2. Schadensvermeidung ..... 26
3. Gerechtigkeit und Fairness ..... 27
4. Autonomie ..... 27

**III. Checkliste 2: Transparenzbericht** ..... 27

1. Abschnitt: Bewertungsphase für die Fragen 2.1. bis 2.6.: Bevor Sie Ihr System entwerfen ..... 27
2. Abschnitt: Bewertungsphase für die Fragen 2.7. bis 2.19: Nach dem Testen des Systems ..... 28
3. Abschnitt: Bewertungsphase für die Frage 2.20: Nach der Implementierung des Systems, wenn das System überwacht wird ..... 29

**IV. Beispiel für den Einsatz der Checklisten 1 und 2: *Swiss COMPAS*** ..... 30

1. Vorbemerkungen ..... 30
2. Checklisten 1 und 2 ..... 31
3. Transparenzbericht ..... 46

/ Flussdiagramm ..... 52

**/ Literaturverzeichnis** ..... 53

**/ Materialienverzeichnis** ..... 61

**Nachwort** ..... 64

# A. Einleitung

Der Einsatz von ADMS hat im öffentlichen Sektor Einzug gehalten. Der von AlgorithmWatch herausgegebene „Automating Society Report", der ein umfassendes Mapping der ADMS-Nutzung in sechzehn europäischen Ländern beinhaltet, zeigt, dass wir bereits heute von einer automatisierten Gesellschaft sprechen können. In den kommenden Jahren wird die Automatisierung von Entscheidungsverfahren und Dienstleistungen voraussichtlich auch in den öffentlichen Verwaltungen exponentiell zunehmen. Die Bevölkerung fordert benutzerfreundliche Dienstleistungen, die einfach und leicht zugänglich sind und rund um die Uhr zur Verfügung stehen. Die Verwaltungen sehen in der Automatisierung eine Chance, die Effizienz zu steigern, Prozesse zu vereinfachen und Routinedienstleistungen zu beschleunigen. Chatbots erleichtern die elektronische Bearbeitung von Steuererklärungen und sparen Ressourcen. Die automatische Bearbeitung von Beschwerden an Behörden beschleunigt Prozesse, die sonst viel Zeit kosten würden. Darüber hinaus werden ADMS eingesetzt, um Anträge auf Sozialleistungen zu bearbeiten, Sozialhilfebetrug aufzudecken, Profile von Arbeitslosen zu erstellen, für die so genannte vorausschauende Polizeiarbeit („Predictive Policing"), oder um das Rückfallrisiko von Bewährungshäftlingen zu bewerten.

Zweifellos bieten ADMS ein großes Potenzial für öffentliche Verwaltungen. Gleichzeitig sind sie aber auch mit erheblichen Risiken verbunden – vor allem, wenn solche Systeme nicht mit der nötigen Sorgfalt eingeführt und eingesetzt werden. Diese Risiken sind nicht auf den öffentlichen Sektor beschränkt (sondern spiegeln oft ähnliche Risiken wider, die den privaten Bereich betreffen), aber sind im öffentlichen Kontext von besonderer Art. Das Handeln der öffentlichen Hand unterliegt nicht nur anderen und einzigartigen rechtlichen Voraussetzungen, wie etwa dem Legalitätsprinzip oder der Einhaltung der Grundrechte. Es findet auch in einem einzigartigen Umfeld statt, in dem Einzelne nicht zwischen verschiedenen Anbieter*innen einer Dienstleistung frei wählen können, sondern unausweichlich einer bestimmten Verwaltung unterworfen ist.

Darüber hinaus haben die Entscheidungen der Behörden oft direkte und schwerwiegende Auswirkungen auf die Einzelnen. Bei der Automatisierung von Verwaltungsverfahren muss dieser besondere Kontext berücksichtigt werden: Es muss in erster Linie darum gehen, dass sie den Menschen und der Gesellschaft nützen und nicht schaden, dass sie die Autonomie der Betroffenen fördern und Grundsätze der Gerechtigkeit und Fairness achten. Nur so handelt die Verwaltung beim Einsatz von ADMS vertrauenswürdig – und nur dann können die Einzelnen und die Bevölkerung als Ganzes der Verwaltung dabei gerechtfertigter Weise vertrauen.

Daher ist es von vorrangiger Bedeutung, den Einsatz von ADMS im öffentlichen Sektor nicht nur zu beobachten und transparent zu machen, sondern auch zu einer grundlegenden ethisch orientierten Reflexion beizutragen: Über die Risiken dieses Einsatzes, seine Auswirkungen auf die Einzelnen und die Gesellschaft sowie die Anforderungen, denen die Systeme genügen müssen. Doch so wertvoll diese Überlegungen auf der Metaebene auch sind, sie reichen nicht aus.

In Anbetracht des einzigartigen Kontextes, in dem Behörden agieren, der besonderen Verantwortung, die sie der Bevölkerung gegenüber wahrnehmen, und der Folgen, die ADMS haben können, sollten die

Behörden verpflichtet sein, die potenziellen Risiken jedes verwendeten Systems umfassend, transparent und systematisch zu bewerten. Um in der Praxis etwas zu bewirken, müssen ethische Grundsätze in praktikable und sofort einsetzbare Instrumente umgesetzt werden, die den Behörden die Mittel an die Hand geben, eine solche Analyse durchzuführen.

In Zusammenarbeit mit der Universität Basel hat AlgorithmWatch im Auftrag des Kantons Zürich eine Studie über den Einsatz von Künstlicher Intelligenz in der öffentlichen Verwaltung durchgeführt. Als Ergebnis dieser Studie haben wir ein konkretes und praktikables Instrument zur Folgenabschätzung entwickelt, das für die Evaluation von spezifischen ADMS durch Behörden auf verschiedenen Ebenen eingesetzt werden kann. Es ermöglicht eine Bewertung auf Einzelfallbasis und während des gesamten Lebenszyklus eines Systems. Damit wollen wir nicht nur einen Beitrag zur Debatte über die ethische Nutzung von ADMS leisten, sondern auch praktische Lösungen anbieten – und damit letztlich sicherstellen, dass diese Systeme tatsächlich zum Nutzen der gesamten Gesellschaft und nicht zu ihrem Nachteil eingesetzt werden.

Wenn es darum geht, die Risiken zu bewerten, brauchen die Beteiligten Transparenz. Ohne Aufklärung über die Funktionsweise konkreter ADMS, ihren Zweck, die beteiligten Akteure und ihre potenziell schädlichen Auswirkungen bleiben sie Black Boxes – für die Verwaltung und ihr Personal, für die Betroffenen und für die Gesellschaft insgesamt. Transparenz garantiert zwar noch nicht die ethische Konformität eines Systems, ist aber eine notwendige Voraussetzung, um diese Konformität zu gewährleisten und Kontrolle und Rechenschaft zu ermöglichen.

Daher stellen wir im Folgenden ein zweistufiges Bewertungsverfahren vor, mit dem die ethisch relevanten potenziellen Implikationen eines bestimmten ADMS erkannt werden können. Es ermöglicht eine Triage solcher Systeme und zeigt auf, ob ein System solche Risikosignale aufweist und zusätzlichen Transparenzanforderungen unterworfen werden muss.

Ist dies der Fall, müssen die Behörden dafür sorgen, dass ein umfassender Transparenzbericht erstellt wird, in dem potentielle Risiken sowie die Maßnahmen, die Behörden zur Risikominderung ergreifen, transparent gemacht werden müssen. Je mehr Risikosignale sich zeigen, desto aufwändiger wird die Erstellung dieses Berichts – und desto anspruchsvoller wird es also für die Verwaltung, das entsprechende System einzusetzen. Damit weicht der hier entwickelte Ansatz von der Vorstellung ab, dass eine Ex-ante-Risikobewertung für bestimmte Kategorien von ADMS pauschal vorgenommen werden könnte. Vielmehr muss jedes ADMS einer einzelfallbezogenen Folgenabschätzung unterzogen werden, die den spezifischen Kontext seines Einsatzes berücksichtigt.

Die Ergebnisse dieser Folgenabschätzung sollten in einem öffentlichen Register offengelegt werden. Solche Register sollten für alle im öffentlichen Sektor eingesetzten ADMS verpflichtend sein und zusätzlich Informationen über den Zweck des Systems, das ihm zugrunde liegende Modell und die an der Entwicklung und Einführung beteiligten Akteure enthalten. Solche Register ermöglichen auch unabhängige externe Überprüfungen, indem sie externen Forscher*innen (Wissenschaft, Zivilgesellschaft und Journalist*innen) Zugang zu relevanten Daten über ADMS geben. Dies trägt zur Forschung im öffentlichen Interesse und damit zu einer faktengestützten Debatte über die Automatisierung des öffentlichen Sektors bei – eine Voraussetzung für die Gewährleistung demokratischer Kontrolle und Rechenschaftspflichten.

Transparenz ist jedoch nicht immer gleichbedeutend mit vollständiger öffentlicher Bekanntgabe. In bestimmten Kontexten können berechtigte Interessen gegen einen vollständigen öffentlichen Zugang zu Transparenzberichten sprechen (wie etwa der Schutz personenbezogener Daten). In solchen Fällen muss jedoch gegenüber bestimmten Stellen, etwa der zuständigen Aufsichtsbehörde, Transparenz hergestellt werden. Diese wiederum muss öffentlich im Register aufgeführt werden.

Zwar müssen ethische Überlegungen zu ADMS letztlich in praktikable Instrumente umgesetzt werden, doch können diese nicht in einem Vakuum entwickelt

werden. Sie müssen auf kohärenten grundlegenden ethischen Analysen und Grundsätzen beruhen, die explizit gemacht werden müssen. Während diese beiden Ebenen in der aktuellen Debatte über ADMS manchmal getrennt werden, halten wir es für entscheidend, dass sie Hand in Hand gehen. Bevor wir also die oben erwähnten Checklisten vorstellen, werden im nächsten Abschnitt die sieben theoretischen Prinzipien, auf denen dieser praktische Werkzeugkasten aufbaut, also die ethische Grundlage unseres Ansatzes, begründet und erläutert.

In Abschnitt D werden die Fragen, die sich aus diesen Grundsätzen ergeben, in zwei separaten Checklisten entwickelt. Die Triage-Checkliste für ADMS (Checkliste 1) hilft bei der Bestimmung, welche ethischen Transparenzfragen vor und während der Einführung eines ADMS behandelt und dokumentiert werden müssen. Die Checkliste zum Transparenzbericht (Checkliste 2) dient dann als Leitfaden für die Erstellung eines detaillierten Transparenzberichts.

Anhand des fiktiven Beispiels eines Schweizer COMPAS-Risikobewertungssystems für Straftäter*innen wird in Abschnitt D.IV die Anwendung des Bewertungsverfahrens illustriert. Ein *Flussdiagramm* gibt schließlich einen Überblick über das gesamte Vorgehen.

# B. Einführung

## I. Ethische Richtlinien für den öffentlichen Sektor

In den vergangenen Jahren gab es einen regelrechten Wettlauf um die Entwicklung von ethischen Richtlinien zum Einsatz KI-basierter Systeme. Weltweit veröffentlichten Unternehmen und Unternehmensverbände, Organisationen der Zivilgesellschaft, Interessen- und Berufsverbände, Regierungen, Behörden und überstaatliche Institutionen Handlungsempfehlungen zum Umgang mit künstlicher Intelligenz. Allein im AI Ethics Guidelines Global Inventory[1], das AlgorithmWatch zusammengestellt hat, sind mehr als 160 solcher Dokumente gesammelt.

Die intensive Debatte um eine «KI-Ethik» hat unterschiedliche Reaktionen ausgelöst. Zum einen wurde das Interesse des Privatsektors misstrauisch beäugt, da Kritiker befürchten, dass dadurch entweder ein soziales Problem in ein technisches verwandelt wird oder Unternehmen versuchen, durch Selbstregulierung strengere Gesetze zu vermeiden. Außerdem wird bisweilen kritisiert, dass Ethikrichtlinien anders als Gesetze nicht demokratisch legitimiert sind. Zum anderen stellen Wissenschaftlerinnen und Wissenschaftler die Frage, ob Einschätzungen dazu, was ein angemessener Einsatz von KI ist und welche Prinzipien die Entwicklung von KI bestimmen werden, konvergieren und sich somit eine Art gemeinsames Verständnis oder gemeinsame Erwartungen an den richtigen Umgang mit KI-basierten Systemen herausbilden.[2]

Im Rahmen dieser Untersuchung treten jene Kritikpunkte, die sich auf den Einsatz von KI-Ethikrichtlinien im Privatsektor beziehen, in den Hintergrund, da der Fokus der Untersuchung auf dem KI-Einsatz in der öffentlichen Verwaltung liegt. Da es bereits eine relevante Anzahl an Handlungsanleitungen gibt, die explizit für die öffentliche Verwaltung als Adressatin verfasst sind, konzentriert sich die vorliegende Studie auf diese Dokumente. Ethikrichtlinien, die sich an alle Entwicklerinnen und Entwickler bzw. alle Anwenderinnen und Anwender richten, werden mithin aus der Untersuchung ausgeklammert.

Dasselbe gilt für sektorspezifische Empfehlungen und Regelungen – unabhängig davon, ob sie sich an Private oder an die öffentliche Verwaltung richten. Gesetze, die Anwendungen von algorithmischen Verfahren regulieren, gibt es bereits seit Langem vom US-amerikanischen *Code of Federal Regulations, Section 255.4 – Display of information*[3] bis hin zur Richtlinie 2014/65/EU des europäischen Parlaments und des Rates vom 15. Mai 2014 über Märkte für Finanzinstrumente sowie zur Änderung der Richtlinien 2002/92/EG und 2011/61/EU.[4] Alle diese gesetzlichen Regulierungen hier zu betrachten, wäre einerseits aus Ressourcengründen nicht möglich. Andererseits wäre dies auch nicht sinnvoll, weil es sich um

1 https://inventory.algorithmwatch.org/about.

2 Jobin/Ienca/Vayena, 2019.

3 «Each [airline reservation] system shall provide to any person upon request the current criteria used in editing and ordering flights for the integrated displays and the weight given to each criterion and the specifications used by the system's programmers in constructing the algorithm.»

4 Hier wird «hochfrequente algorithmische Handelstechnik» als algorithmische Handelstechnik definiert, die u. a. «gekennzeichnet ist durch [...] b) die Entscheidung des Systems über die Einleitung, das Erzeugen, das Weiterleiten oder die Ausführung eines Auftrags ohne menschliche Intervention».

sektorspezifische Regelungen handelt, die wenig zur Beantwortung der Frage beitragen können, welche generellen Aspekte die öffentliche Verwaltung beim Einsatz von KI-basierten bzw. automatisierten Entscheidungssystemen beachten sollte. Deshalb werden die genannten Richtlinien und Regelungen hier grundsätzlich ausgeklammert.

Im Einzelnen werden die folgenden Richtlinien einbezogen:

**Supranationale Richtlinien/verschiedene Akteure:**

1. Article 29 Data Protection Working Party
   *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*[5]

2. Europarat
   *Discrimination, artificial intelligence, and algorithmic decision-making* (insbesondere S. 29: Public sector bodies)[6]

3. Europarat
   *Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems* (insbesondere Anhang A, Nr. 11 und 12)[7]

4. Europarat
   *European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment*[8]

5. Dataethics
   *Data Ethics in Public Procurement*[9]

6. World Economic Forum
   *Unlocking Public Sector AI: AI Procurement in a Box*[10]

7. Cities Coalition for Digital Rights
   *Transparency, accountability, and non-discrimination of data, content and algorithms*[11]

8. AI Now Institute, City of Amsterdam, City of Helsinki, Mozilla Foundation, Nesta
   *Using procurement instruments to ensure trustworthy AI*[12]

9. ePaństwo Foundation
   *alGOVrithms: The Usage of Automated Decision Making – Policy Recommendations For Decision Makers*[13]

**Nationale Richtlinien**

10. Australien: Commonwealth Ombudsman
    *Automated decision-making better practice guide*[14]

11. Kanada: Government of Canada
    *Directive on Automated Decision-Making*[15]

12. Deutschland: Kompetenzzentrum Öffentliche IT
    *KI im Behördeneinsatz: Erfahrungen und Empfehlungen*[16]

5 https://ec.europa.eu/newsroom/article29/document.cfm?action=display&doc_id=49826.
6 https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73.
7 https://rm.coe.int/09000016809e1154.
8 https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c.
9 https://dataethics.eu/publicprocurement/.
10 http://www3.weforum.org/docs/WEF_AI_Procurement_in_a_Box_Project_Overview_2020.pdf.
11 https://citiesfordigitalrights.org/#declaration.
12 https://assets.mofoprod.net/network/documents/Using_procurement_instruments_to_ensure_trustworthy_AI.pdf.
13 https://idfi.ge/public/upload/IDFI_2019/General/alGOVrithms%20-%20Recommendations%20EN.pdf.
14 https://www.lawcouncil.asn.au/publicassets/afebc52d-afa6-e911-93fe-005056be13b5/3639%20-%20AI%20ethics.pdf.
15 https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592.
16 https://www.oeffentliche-it.de/documents/10181/14412/KI+im+Beh%C3%B6rdeneinsatz+-+Erfahrungen+und+Empfehlungen

13. Neuseeland: Government of New Zealand
*Algorithm charter for Aotearoa New Zealand*[17]

14. Schweiz: Bundesrat
*Leitlinien «Künstliche Intelligenz» für den Bund*[18]

15. Großbritannien: Government Digital Service and Office for Artificial Intelligence UK
*A guide to using artificial intelligence in the public sector*[19]

16. Großbritannien: National Health Service
*Code of conduct for data-driven health and care technology*[20]

17. Vereinigte Staaten von Amerika: New York City
*Automated Decision Systems Task Force Report*[21]

## II. Zweistufiges Beurteilungsverfahren

Die erwähnten Richtlinien schaffen zwar Eckpunkte und geben Hinweise für einen ethisch vertretbaren Einsatz von KI, sind für die konkrete Umsetzung von KI-Vorhaben in der öffentlichen Verwaltung aber nicht ohne weitere Implementierungsschritte umsetzbar. In diesem Kapitel wird deshalb ein zweistufiges Beurteilungsverfahren entwickelt, das dazu genutzt werden kann, ethische Auswirkungen eines KI-Systems zu erkennen und darauf aufbauend Transparenz über das System herzustellen.

Im Folgenden wird zunächst dargelegt, warum in diesem Bericht zum einen bestimmte *ethische Grundsätze* und zum anderen bestimmte *instrumentelle Grundsätze* als Grundlage einer Beurteilung von KI-Systemen herangezogen werden. Anschließend wird in den Abschnitten B.I. und B.II. zu jedem der sieben identifizierten Grundsätze detailliert dargestellt, welche Fragen aus diesen Grundsätzen folgen und bei der Beurteilung entsprechend zu beantworten sind. Die Fragen selbst werden in zwei verschiedenen Checklisten in Abschnitt C dargestellt. Die *Triage-Checkliste für KI-Systeme* (Checkliste 1) hilft bei der Feststellung, welche ethischen Transparenzfragen vor und während der Durchführung eines KI-Projektes im Detail zu dokumentieren sind. Die *Checkliste Transparenzbericht* (Checkliste 2) dient als Leitfaden für die Erstellung eines ausführlichen Transparenzberichts. In den folgenden Ausführungen wird auf Fragen aus den beiden Checklisten in den Abschnitten D.II. und D.III. jeweils mittels kursiver Angabe (z. B. *Frage. 1.10* oder *Frage 2.8*) verwiesen. Wie das skizzierte Beurteilungsverfahren angewendet werden kann, wird sodann am fiktiven Beispiel eines *Swiss-COMPAS*-Risikobewertungssystems für Straftäterinnen und Straftäter veranschaulicht (Abschnitt D.IV.). Ein *Flussdiagramm* gibt schließlich einen Überblick über das gesamte Vorgehen.

Als Grundlage des hier entwickelten Beurteilungsverfahrens dienen die «Ethik-Leitlinien für eine vertrauenswürdige KI» der hochrangigen Expertengruppe für Künstliche Intelligenz, die von der Europäischen Kommission eingesetzt wurde.[22] Dieses Dokument stellt allerdings lediglich eine Vereinfachung derjenigen Grundwerte dar, die in anderen Richtlinien erarbeitet wurden, und muss daher ergänzt werden.

Zu diesem Zweck werden weitere Zusammenfassungen von Richtlinien herangezogen. Die umfassendste Analyse bisher veröffentlichter KI-Richtlinien[23] enthält eine Liste von elf verschiedenen Grundsätzen, die in den analysierten Richtlinien als gemeinsamer Nenner enthalten sind. Von diesen elf Grundsätzen über-

17 https://data.govt.nz/use-data/data-ethics/government-algorithm-transparency-and-accountability/algorithm-charter.
18 https://www.sbfi.admin.ch/dam/sbfi/de/dokumente/2020/11/leitlinie_ki.pdf.download.pdf/Leitlinien%20K%C3%BCnstliche%20Intelligenz%20-%20DE.pdf.
19 https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector.
20 https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology.
21 https://www1.nyc.gov/assets/adstaskforce/downloads/pdf/ADS-Report-11192019.pdf.
22 Ethics Guidelines for Trustworthy AI, 2019.
23 Jobin/Ienca/Vayena, 2019.

schneiden sich einige (**Nicht-Missbräuchlichkeit oder Schadensverhütung, Gerechtigkeit und Unparteilichkeit [Fairness] sowie Freiheit und Autonomie**) mit den ethischen Grundsätzen der Richtlinien der Expertengruppe. Zwei der elf *ethischen Grundsätze* sind *so* nicht in den EU-Leitlinien enthalten: Benefizienz und Achtung der Würde. In den elf Grundsätzen sind zudem instrumentelle, technische oder verfahrenstechnische Anforderungen (**Transparenz** und **Verantwortung/Rechenschaftspflicht**) enthalten, die in den EU-Richtlinien als «Schlüsselanforderungen für vertrauenswürdige KI» bezeichnet werden. Ein weiterer Grundsatz, die **Erklärbarkeit**, ist in den EU-Leitlinien ebenfalls enthalten, wird aber – was auch plausibel ist – als Grundlage für andere Grundsätze wie die Implementierungsanforderungen betrachtet.[24] Der Grundsatz der **Benefizienz** ist nicht in den EU-Leitlinien enthalten, aber sie ist nicht nur in der erwähnten Zusammenfassung von *Jobin et al.*[25] aufgeführt, sondern auch im am weitesten verbreiteten Grundgerüst ethischer Grundsätze, jenem der biomedizinischen Ethik, enthalten.[26]

Als weitere Grundsätze werden als Makrokategorien die **Kontrolle**, die **Transparenz** und die **Rechenschaftspflicht** betrachtet. *Loi et al.* verstehen darunter Maßnahmen in den KI-Ethikrichtlinien, die *instrumentelle und verfahrenstechnische Anforderungen* umfassen.[27] Im Folgenden werden daher Kontrolle, Transparenz und Rechenschaftspflicht als «*instrumentelle Grundsätze*» bezeichnet.[28]

Die Analyse von 18 weiteren Dokumenten zum Einsatz von KI im öffentlichen Sektor (siehe A.I.) ergibt ebenfalls ethische und instrumentelle Grundsätze, die mit diesem Gerüst vereinbar sind.

Im Folgenden konzentriert sich die Analyse daher auf einen Rahmen von sieben Werten:

— drei der vier ethischen Grundsätze, die in den EU-Leitlinien enthalten sind, d. h. die Achtung der **menschlichen Autonomie,** die **Schadensvermeidung sowie** die **Gerechtigkeit oder Unparteilichkeit** [Fairness];

— die **Benefizienz** als weithin anerkannter ethischer Grundsatz sowie

— die drei instrumentellen Grundsätze **der Kontrolle, Transparenz und Rechenschaftspflicht**, die technische, organisatorische und aufsichtsrechtliche Anforderungen zusammenfassen, die üblicherweise in praktischen Richtlinien zur KI-Ethik enthalten sind.

Im Gegensatz zu den Richtlinien der EU-Expertengruppe wird die *Erklärbarkeit* nicht als eigenständiger Grundsatz, sondern als Bestandteil anderer instrumenteller Grundsätze betrachtet.[29]

Im Folgenden wird der ethische Rahmen, der die Grundlage der in dieser Untersuchung entwickelten praktischen Empfehlungen und Checklisten bildet, näher ausgeführt. Für jeden der sieben berücksichtigten Werte wird in Fußnoten auf analoge Konzepte in bestehenden Richtlinien für Anwendungen von KI-Systemen im öffentlichen Sektor verwiesen. Darüber hinaus wird auf die Fragen in den Checklisten 1 und 2 verwiesen, die direkt aus dieser Analyse abgeleitet werden.

Für die Schweiz sind die Leitlinien des Schweizer Bundesrates zu KI von besonderem Interesse. Die darin enthaltene Idee, «den Menschen in den Mittelpunkt zu stellen», findet sich in fünf ethischen Grundsätzen wieder, die hier ebenfalls berücksichtigt werden: die Benefizienz als Förderung des Wohlergehens, die Gerechtigkeit als Achtung der Grundrechte und

24 Floridi/Cowls, 2019.
25 Jobin/Ienca/Vayena, 2019.
26 Beauchamp/Childress, 2008.
27 Loi/Heitz/Christen, 2020.
28 Loi, 2020.
29 Vgl. auch Jobin/Ienca/Vayena, 2019 und Loi/Heitz/Christen, 2020.

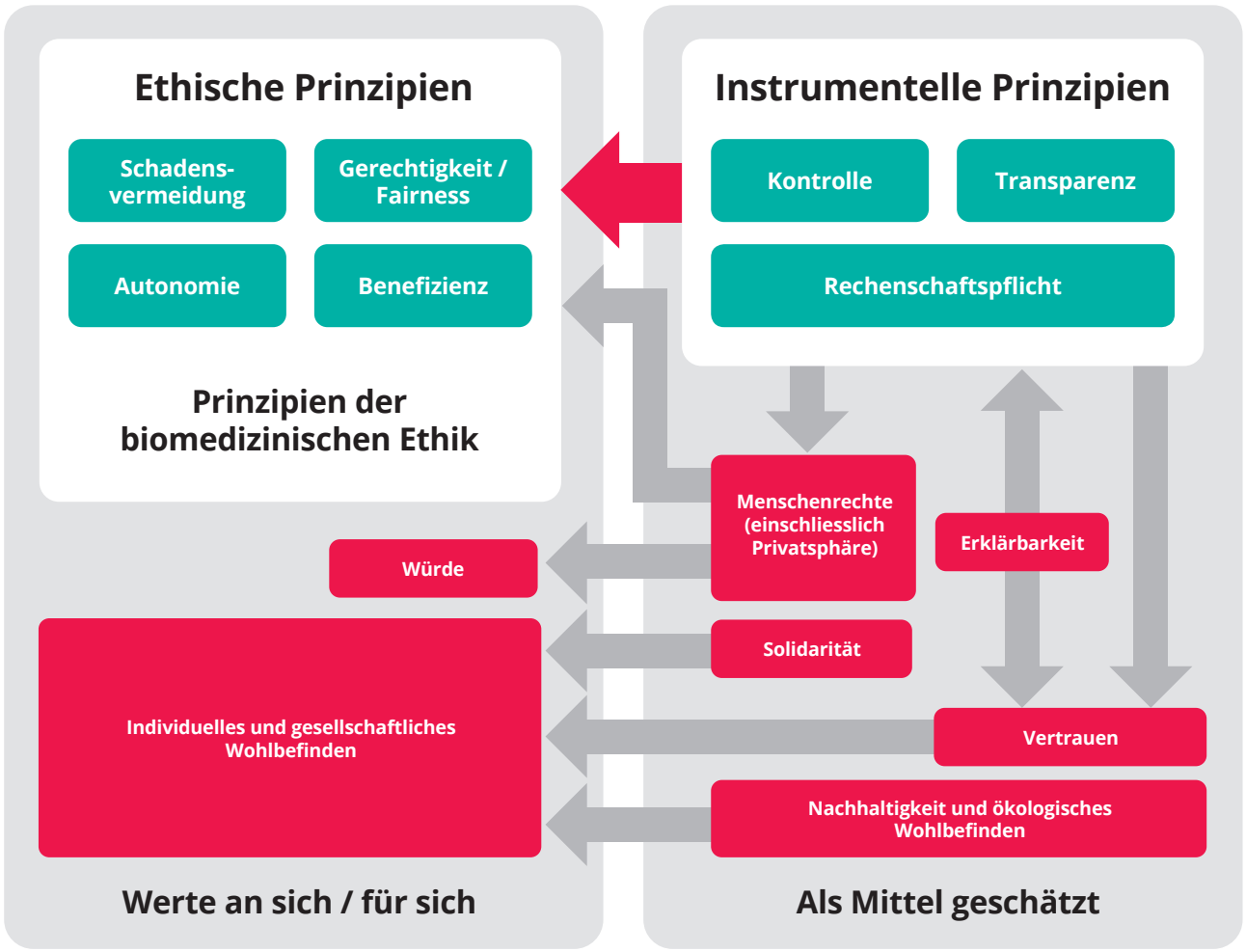



Abb. 1 (eigene Darstellung): Die wichtigsten Grundsätze und Werte in den ethischen Richtlinien zur KI. Grün eingefärbt sind die in dieser Untersuchung berücksichtigten Grundsätze, rot sind Werte und Grundsätze aus anderen Richtlinien. Der Pfeil bedeutet «ist erforderlich für».

Vermeidung von Diskriminierung, die Autonomie als Selbstbestimmung, die Schadensvermeidung als Schutz der Privatsphäre und personenbezogener Daten sowie die Würde.[30] Somit kann festgehalten werden, dass die folgenden Erwägungen mit den KI-Leitlinien des Bundesrates grundsätzlich kompatibel sind.

30 Bundesrat Leitlinien KI 2020, Leitlinie.

# C. Sieben Grundsätze

## I. Ethische Grundsätze

### 1. Schadensvermeidung

Dies ist das Prinzip «Niemandem einen Schaden zufügen». Zivile KI-Systeme dürfen nicht so konzipiert sein, dass sie Menschen schaden oder täuschen, und sie sollten so implementiert werden, dass negative Ergebnisse minimiert werden.[31] Schaden zu vermeiden, bedeutet in erster Linie in der einfachsten Form, Schmerzen und Unbehagen für Menschen zu verhindern. Im weiteren Sinne umfasst Schaden die Verletzung der Privatsphäre[32] (*Frage 1.1*) und der Rechte (*Fragen 1.4 und 1.5*) einschließlich der Menschenrechte.[33] Die Vermeidung von Schäden ist damit verbunden, Sicherheit[34] und Nachhaltigkeit zu fördern sowie allgemeine, technische und institutionelle Schutzmaßnahmen aufzubauen.[35] Diese werden oft als Vertrauenselement oder vertrauenswürdige Technologie bezeichnet.[36] Zur Vermeidung von Schäden gehören auch die Verhinderung und Steuerung von Risiken. Häufig erfordert das, zu gewährleisten, dass die Systeme zuverlässig[37] und berechenbar[38] sind. Schadensvermeidung umfasst ebenfalls, eine breite ökologische und soziale Nachhaltigkeit sicherzustellen (*Frage 1.10*)[39], denn nicht nachhaltige Praktiken führen letztendlich zu menschlichem Schaden und zu einer Verletzung eigener Interessen. Darüber hinaus kann hier die Nachhaltigkeit eines *sozio-technischen Systems* berücksichtigt werden, das auf einer zuverlässigen Technologie basiert.[40] Dies umfasst auch die Gewährleistung der Cybersicherheit (*Frage 1.2*) einschließlich des Schutzes der Vertraulichkeit, Integrität und Verfügbarkeit von Informationen.[41]

### 2. Gerechtigkeit und Fairness

Das ethische Ziel von Gerechtigkeit und Fairness umfasst die Wahrung von sechs Dimensionen ethischer Werte. Die erste (in Bezug darauf, wie oft sie in den KI-Ethikrichtlinien erwähnt wird) ist der Schutz vor einer *ungerechten Diskriminierung* und *nicht zu rechtfertigenden Voreingenommenheit*.[42] Diese Dimen-

31 Dawson et al., 2020, Principle 2.
32 Dawson et al., 2020, Principle 4.
33 Council of Europe, CM/Rec (2020)1, Principle 3.2; Government of Canada, 2019, Appendix B (risk evaluation framework).
34 Council of Europe, CM/Rec (2020)1; Bundesrat Leitlinien KI 2020, Leitlinie 5; Engelmann/Puntschuh, 2020, 5. Sicherheit.
35 Council of Europe, CM/Rec (2020)1; Engelmann/Puntschuh, 2020, 5. Sicherheit.
36 Leslie, 2019; Ethics Guidelines for Trustworthy AI, 2019.
37 Government of New Zealand, 2020.
38 Vgl. Government of Canada, 2019, Guidelines and laws that require a risk assessment may include: "the rights of individuals or communities, the health or well-being of individuals or communities, the economic interests of individuals, entities, or communities, the ongoing sustainability of an ecosystem".
39 Council of Europe, CM/Rec (2020)1, Principle 6.3.
40 Government UK, 2019.
41 Council of Europe, 2020, Ethical Charter AI, Principle of quality and security.; Dawson et al., 2020, Principle 4; Government of Canada, 2019, 6.3.7; Engelmann/Puntschuh, 2020, 6. Datenhaltung und -qualität.
42 Council of Europe, 2020, Ethical Charter AI, Principle of non-discrimination; Dawson et al., 2020, Principle 5; Government of New Zealand, 2020, Fairness and Justice; Leslie, 2019; New York City, 2019.

sion der Fairness gilt für verschiedene Elemente der Datenverarbeitung:

Wenn das KI-System soziale oder demografische Daten verarbeitet, sollte es so ausgelegt werden, dass ein Mindestmaß an Schadensvermeidung in Bezug auf Diskriminierung erreicht wird. Um dies zu tun,

— sollten nur faire und gerechte Datensätze verwendet werden (Datengerechtigkeit),

— sollten angemessene Funktionen, Prozesse und analytische Strukturen in die Modellarchitektur aufgenommen werden (Design-Fairness),

— sollte verhindert werden, dass das System diskriminierende Auswirkungen hat (Ergebnisgerechtigkeit), sowie

— das System unvoreingenommen implementiert werden (Implementierungsgerechtigkeit).[43]

Wahrscheinlich ist es nicht möglich, jeden *Bias* zu vermeiden. Voreingenommenheit kann auf unzureichende Daten beim Trainieren des Modells zurückzuführen sein,[44] aber es kann sich schwierig oder unmöglich gestalten, repräsentative Daten zum Trainieren von Modellen zu finden – auch aufgrund von Datenschutzbestimmungen und (paradoxerweise) Antidiskriminierungsvorgaben, die es erschweren, beispielsweise Daten über sexuelle Orientierung, Geschlechter, ethnische Gruppen oder religiösen Status zu sammeln und zu verarbeiten. Aber selbst dann, wenn die Daten völlig ausreichen, um die Gesellschaft in allen ihren Facetten darzustellen, können Entscheidungen, die auf statistischen Verallgemeinerungen beruhen, als ungerechtfertigt voreingenommen angesehen werden – etwa, wenn sie für Personen mit Merkmalen, die mit denen aus privilegierteren Umständen übereinstimmen, günstiger sind. Das ist *insbesondere* dann der Fall, wenn Daten vorhanden sind, die für verschiedene gesellschaftliche Gruppen repräsentativ sind, da dies die Wahrscheinlichkeit erhöht, dass einige Daten (oder eine Kombination davon) als Stellvertreter für Alter, Geschlecht, Religion usw. fungieren. Auch wenn solche Daten, die *explizit* die Kategorien betreffen, die durch Antidiskriminierungsgesetze geschützt sind, nicht in dem Mix enthalten sind, wird jedes effiziente Verfahren maschinellen Lernens, das Algorithmen erzeugt, normalerweise lernen, Stellvertreter zu erkennen. Aus diesem Grund sind alle algorithmischen Schlussfolgerungen, die auf Techniken des statistischen Lernens aus Daten über Menschen basieren, in Bezug auf *Bias* und indirekte Diskriminierung potenziell moralisch problematisch. Daher muss diesen Algorithmen in der Checkliste (*Frage 1.13*) besondere Aufmerksamkeit gewidmet werden.

Darüber hinaus ist die Dimension der Ergebnis- und Anwendungsgerechtigkeit besonders wichtig, wenn Algorithmen den Wettbewerb in Politik und Wirtschaft beeinflussen (*Fragen 1.11* und *1.12*). Wettbewerbsprozesse sind von enormer Bedeutung, da sich die Gesellschaft auf einen fairen Wettbewerb in Markt und Politik stützt, um zu entscheiden, wie soziale Ressourcen und Chancen in der Gesellschaft auf eine Weise verteilt werden müssen, die insgesamt als verfahrensgerecht betrachtet werden kann.[45]

Zweitens gehen die Fairnessanforderungen über die Ethik hinaus und beziehen die *Legalität* mit ein, um sicherzustellen, dass Algorithmen nicht gegen bestehende Gesetze einschließlich der gesetzlichen Rechte verstossen.[46]

Drittens ist die *Achtung aller Rechte*, ob sie im positiven Recht anerkannt sind oder nicht, einschließlich der Menschenrechte und Persönlichkeitsrechte beinhaltet.[47]

43 Government UK, 2019.
44 Council of Europe, CM/Rec (2020)1, Testing on personal data.
45 Rawls, 1999.
46 Dawson et al., 2020, Principle 3; Government of New Zealand, 2020, Fairness and Justice.
47 Government of New Zealand, 2020, Fairness and Justice.

Viertens ist der Wert von *Gleichheit,*[48] *Inklusion und Solidarität* erfasst (obwohl dies ein umstrittenerer Wertekanon sein kann und sehr kontextabhängig ist). Das Erfordernis der Inklusion, das Elemente der öffentlichen Beteiligung umfasst, scheint in Leitlinien für den öffentlichen Sektor stärker vertreten zu werden[49] als in Leitlinien für die Wissenschaft oder für private Unternehmen, die zumeist früher entwickelt worden sind.

Fünftens sind *Entschädigungen und Rechtsmittel* eingeschlossen (*Fragen 1.7, 1.8, 1.9*), wenn eine Rechtsverletzung nachgewiesen werden kann.[50]

Sechstens geht es um die Frage der *prozessualen Ordnungsmäßigkeit*, die in den untersuchten Leitlinien nicht erwähnt wurde, aber in der Literatur berücksichtigt wird. Einige KI-Systeme aktualisieren ihre Modelle kontinuierlich basierend auf neuen Daten, um das Ziel, für das sie programmiert wurden, besser zu erreichen. Ein Nebeneffekt eines kontinuierlich aktualisierten Modells besteht darin, dass es unterschiedliche Ergebnisse für dieselben Eingaben erzeugen kann, wenn dieselben Eingaben vor oder nach einer Modellaktualisierung verarbeitet werden. Dies bedeutet, dass zwei Personen mit denselben Merkmalen (Eingaben) möglicherweise unterschiedliche Entscheidungen (Ausgaben) erhalten, die davon abhängen, wann der Algorithmus ihre personenbezogenen Daten verarbeitet (vor oder nach einer Modellaktualisierung).[51] Dies kann das Recht der Personen auf rechtsgleiche Behandlung verletzen (*Frage 1.14*).

## 3. Autonomie

Die Förderung der Autonomie bedeutet, dass Einzelne Entscheidungen über ihr Leben treffen können, die ihre eigenen sind und die ihnen nicht von anderen auferlegt oder von ihnen manipuliert werden. Eine Entscheidung auf der Grundlage unzureichender Informationen oder einer Täuschung gilt nicht als autonom. Das Ziel der menschlichen Autonomie hängt hauptsächlich mit dem prozessualen Erfordernis der Transparenz zusammen, das im zweiten Teil der Leitlinie beschrieben wird. Transparenz bedeutet, ausreichende Informationen bereitzustellen und die Täuschung von Personen zu vermeiden, die mit den Algorithmen interagieren, was autonome Entscheidungen ermöglicht.

Die bekannteste (und vielleicht am wenigsten geschätzte) Implementierung von Entscheidungsautonomie im digitalen Alltag betrifft personenbezogene Daten. Individuen müssen darüber informiert werden, was mit ihren persönlichen Daten geschehen kann, damit sie ihre Zustimmung dazu geben können, dass ihre Daten auf die eine oder andere Weise verwendet werden.[52] Dies gilt insbesondere im Zusammenhang mit experimentellen Technologien.[53]

Ein weiterer Aspekt der Autonomie ist die «Fähigkeit, unfaire, voreingenommene oder diskriminierende Systeme infrage zu stellen und zu ändern»,[54] über die Bürgerinnen und Bürger nur verfügen, wenn sie «verständliche und genaue Informationen über die technologischen, algorithmischen und Künstlichen Intelligenzsysteme erhalten, die sich auf ihr Leben auswirken».[55] Die Anfechtung kann *sowohl* als wertvoll *an sich* als Element menschlicher Autonomie *als auch* als instrumentell wertvoll als eine Form der Kontrolle des Algorithmus und ein Weg zur Förderung der

48 Government of New Zealand, 2020, Fairness and Justice.

49 Council of Europe, CM/Rec (2020)1, Barriers, Advancement of public benefits; Dataethical Thinkdotank, 2021, Universal design; Cities for Digital Rights, 2020, Participatory democracy, diversity and inclusion; Bundesrat Leitlinien KI 2020, Leitlinie 7.

50 Council of Europe, CM/Rec (2020)1, Effective remedies; Government of Canada, 2019, Principle 6.4.

51 Kroll et al. 2016/2017; Loi/Ferrario/Viganò, 2020.

52 Cities for Digital Rights, 2020, Privacy, data protection and security; WEF, 2020.

53 Council of Europe, CM/Rec (2020)1, computational experimentation.

54 Cities for Digital Rights, 2020, Transparency, accountability, and non-discrimination of data, content and algorithms.

55 Cities for Digital Rights, 2020, Transparency, accountability, and non-discrimination of data, content and algorithms.

Rechenschaftspflicht angesehen werden (Letzteres wird in Teil 2 dieser Leitlinien betrachtet).

Ein weiterer Aspekt der Autonomie ist die Möglichkeit, die zu verwendenden digitalen Dienste auszuwählen oder deren Einsatz ganz zu vermeiden,[56] insbesondere dann, wenn es sich um experimentelle Dienste handelt[57] (*Frage 1.6*).

Autonomie hat auch eine kollektive Dimension: Sie ist die Fähigkeit der Bürger, gemeinsam Entscheidungen über ihr kollektives Schicksal als Gemeinschaft zu treffen. Diese kollektive Dimension der Autonomie ist in KI-Richtlinien nicht weit verbreitet,[58] scheint jedoch für öffentliche digitale Infrastrukturen wie die von Smart Cities sehr wichtig zu sein.[59]

Das ethische Erfordernis, die Grundrechte zu respektieren,[60] fördert implizit die Autonomie, da viele dieser Rechte (die typischerweise die Menschenrechte respektieren und in Demokratien verfassungsrechtlich geschützt sind) die Autonomie des Menschen schützen. Beispielsweise gewähren negative Rechte wie die Meinungs- oder Religionsfreiheit der individuellen Autonomie im politischen Raum und bei individueller und kollektiver Meinungsäußerung Schutz. Positive Rechte wie das Recht auf Gesundheitsversorgung und Bildung schützen die Autonomie, indem sie sicherstellen, dass der Einzelne über die Mittel verfügt, die er für ein unabhängiges Leben benötigt. Daher wird die Autonomie durch eine sozial nachhaltige KI gefördert.[61]

Schließlich kann sich Autonomie in der KI-Ethik auf die Idee beziehen, dass das KI-System «unter Benutzerkontrolle» steht.[62] Das bedeutet, Algorithmen sollten verwendet werden, um die menschliche Entscheidungsfindung zu unterstützen, und nicht dazu dienen, sie vollständig zu ersetzen. Dies wird am plausibelsten als ein eingeschränktes Prinzip verstanden. Philosophisch gesehen kann man argumentieren, Automatisierung könne die Autonomie eher *erweitern* als reduzieren (und sie hat es auch getan), insofern die Automatisierung von Routineaufgaben dazu beigetragen hat, mehr Zeit und Ressourcen für den Menschen freizugeben, damit er sich mit intellektuell herausfordernden, kreativen oder emotional lohnenden Aufgaben befassen kann.[63] Das Problem der Autonomie ergibt sich aus KI-Systemen, die kognitiv anspruchsvollere Arten menschlicher Aktivitäten automatisieren sollen, wobei Menschen Befehle *von* den Maschinen entgegennehmen, anstatt ihnen Befehle zu erteilen.[64] Auf dem Spiel steht Autonomie als ethischer Wert daher möglicherweise bei all jenen Automatisierungsprojekten, bei denen KI-Systeme das menschliche Urteilsvermögen ersetzen sollen (*Frage 1.15*), und bei den Formen der Automatisierung, bei denen unklar ist, ob Benutzerinnen und Benutzer die KI-Entscheidungen ausreichend verstehen, damit diese ihre Autonomie *unterstützen*, statt sie durch KI-Systeme zu *ersetzen* (*Fragen 1.16* und *1.17*). Darüber hinaus kann die Öffentlichkeit die Kontrolle und damit die Autonomie über ihre Prozesse und Entscheidungen verlieren, wenn sie sich auf eine Infrastruktur stützen, die sich vollständig im Besitz von Dritten befindet und von ihnen abgeschottet wird (*Frage 1.18*). Dies ist ein aufkommendes Problem, das in früheren Überprüfungen nicht erkennbar war,[65] aber in Bezug auf die KI-Systeme des öffentlichen Sektors ziemlich wichtig erscheint.[66]

---

56 Cities for Digital Rights, 2020, Open and ethical digital service standards.

57 Council of Europe, CM/Rec (2020)1, Computational experimentation.

58 Es wurde nur von Jobin et al., 2019, erwähnt.

59 Vgl. Cities for Digital Rights, 2020: "Everyone should have [...] the ability collectively to engage with the city through open, participatory and transparent digital processes", Participatory Democracy, diversity and inclusion.

60 Council of Europe, CM/Rec (2020)1, Principle of respect for fundamental rights.

61 Council of Europe, CM/Rec (2020)1, Human-centric and sustainable innovation; Dataethical Thinkdotank, 2021, Human primacy.

62 Council of Europe, CM/Rec (2020)1, Principle under user control.

63 Danaher, 2016a, Danaher, 2016b; Loi, 2015.

64 Danaher, 2016a, Danaher, 2016b; Loi, 2015.

65 Jobin/Ienca/Vayena, 2019.

66 Cities for Digital Rights, 2020, Participatory Democracy, diversity and inclusion and Open and ethical digital service standards; Council of Europe, CM/Rec (2020)1, Infrastructure.

### 4. Benefizienz

Benefizienz ist wohl dasjenige Grundprinzip der Ethik, das in den KI-Richtlinien am wenigsten verbreitet ist. Ein plausibler Grund für die unzulängliche Beachtung der Benefizienz ist, dass die meisten Akteurinnen und Akteure, die sich mit KI-Systemen befassen, davon ausgehen, KI-Systeme könnten irgendwelche Vorteile bringen. Effizienz wird häufig als Grund für die Verwendung von KI genannt: Der gleiche Dienst kann für die gleiche Anzahl von Personen bereitgestellt werden, während weniger Ressourcen verwendet oder vorhandene Dienste können verbessert werden (z. B., indem sie genauere Ergebnisse liefern oder mit zusätzlichen Funktionen ausgestattet werden), während sie günstig und damit für die meisten zugänglich bleiben. Und doch ist die Möglichkeit, mithilfe von KI-Systemen *Gutes zu tun*, ethisch grundlegend, da eine Leitlinie für die ethische Verwendung von KI-Systemen zu stark auf Schadensverhütung und zu wenig auf die Schaffung von Nutzen ausgerichtet sein kann. Eine solche Leitlinie wird sich in den meisten Kontexten eher gegen die Einführung von KI-Systemen aussprechen, da Innovation an sich Risiken birgt. Wenn man den potenziellen Nutzen von Innovation vergisst, gibt es keinen Grund, *irgendein Risiko* einzugehen. Eine extreme *Risikovermeidung* ist jedoch nicht immer sinnvoll. Vielmehr sollte das mit Innovationen verbundene Risiko *gemanagt* werden. Beispielsweise haben KI-Systeme wie erwähnt das Potenzial, die Autonomie des Menschen zu verbessern, wenn sie dazu verwendet werden, Prozesse so zu automatisieren, dass menschliche Ressourcen freigesetzt werden, die besser an anderer Stelle eingesetzt werden. Glücklicherweise erwähnen einige Richtlinien, die sich an den öffentlichen Sektor richten, zumindest implizit die Benefizienz, indem sie auf den Nutzen von Innovationen hinweisen.[67] Zum Beispiel: «Erzeugt Nettonutzen. Das KI-System muss Vorteile für Menschen generieren, die mehr wiegen als die Kosten.»[68] Die Benefizienz wird implizit auch in diesen Leitlinien angesprochen, in denen die Förderung des menschlichen Wohlbefindens als übergeordnetes Ziel einer solchen Innovation angegeben ist.[69] Die Richtlinien für den öffentlichen Sektor betonen den Gedanken, dass der Nutzen der KI-Systeme ein *öffentlicher* sein sollte.[70]

Die Benefizienz wird in der Checkliste nicht ausdrücklich aufgeführt. Das Konzept eines Nettonutzens wird jedoch implizit im Transparenzbericht (Checkliste 2) erwähnt, insbesondere in den *Fragen 2.1, 2.16* und *2.17*, in denen erläutert werden muss, warum die Einführung von KI-Systemen *nützlich* ist und welche Nachweise dafür erbracht werden können. Darüber hinaus führt die Frage nach dem Schaden für das Gemeinwohl (*Frage 1.10*) (gemäß dem vorgeschlagenen Checklistenalgorithmus) zu einer transparenten Stakeholderanalyse (*Frage 2.8*) und einer transparenten Darstellung möglicher Kritik, die externe Stakeholder (*Frage 2.20*) gegen die Kosten-Nutzen-Analyse der öffentlichen Verwaltung vorbringen.

## II. Instrumentelle und aufsichtsrechtliche Grundsätze

### 1. Kontrolle

Das prozessuale Erfordernis der Kontrolle ergibt sich aus der Analyse von 20 Richtlinien, die sich eher auf Handlungstypen als auf Handlungsziele konzentrierten.[71] In den ursprünglich analysierten Leitlinien erschien dieses Erfordernis nicht als eigenständige Anforderung, sondern eher als ein Element – eine gemeinsame Reihe von Handlungen –, das jeweils gleich häufig unter den Rubriken *Transparenz* und *Rechenschaftspflicht* aufgeführt ist. In der Tat umfasst die Kontrolle gemeinsame Aktivitäten, die sowohl für die Transparenz als auch für die Rechenschaftspflicht erforderlich sind: Man kann in Bezug auf

67 Engelmann/Puntschuh, 2020, 2. Interne Veränderung und 3. Innovationsmarker.
68 Dawson et al., 2020.
69 Leslie, 2019; Government of New Zealand, 2020, Well-Being.
70 Council of Europe, CM/Rec (2020)1, Advancement of public benefit, Rights-promoting technology.
71 Loi/Heitz/Christen, 2020.

Prozesse oder Ergebnisse, die man nicht kennt, nicht transparent sein und Verantwortung im positiven, vorausschauenden Sinne des Begriffs (nicht im Sinne von rückwirkender Schuld oder Haftung) umfasst, Prozesse so zu kontrollieren, dass sie zu den beabsichtigten Ergebnissen führen. Die Kontrolle umfasst alle Aktivitäten, die erforderlich sind, um allen zielbezogenen Aktivitäten *Robustheit* zu verleihen: vom Erreichen des von den Benutzerinnen und Benutzern vorgesehenen Ziels des KI-Systems (welches immer das sein mag) bis hin zur Gewährleistung, dass die anderen ethischen Ziele (Schadensvermeidung, Fairness und Autonomie) ebenfalls gefördert werden. Kontrolle erscheint moralisch neutral, weil ihr ethischer Wert rein *instrumentell* und *ungewiss* ist. Eine Reihe von Praktiken, mit denen Terroristen ein KI-System besser kontrollieren können, um Drohnenangriffe zu koordinieren, ist nützlich, aber nicht ethisch wertvoll, da das Ziel, das Terroristen verfolgen, als solches schlecht ist. Wenn das Ziel des oder der Handelnden jedoch ethisch positiv ist, ist das Fehlen von Kontrolle ethisch schlecht und nicht nur neutral, da gute Absichten ohne Kontrolle möglicherweise nicht das Gute erreichen, auf das sie abzielen, oder sogar unbeabsichtigt schädlich sein können.

Da es sich bei Kontrolle um ein so vielseitiges Mittel handelt, ist es nicht überraschend, dass sie die häufigste Verfahrensanforderung ist, die in den Ethikrichtlinien für KI-Systeme enthalten ist.

Erstens umfasst die Kontrolle die Dokumentation von Prozessen und Ergebnissen sowie die *Aufzeichnung*,[72] *Prüfung*[73] und *Überwachung*,[74] die die Daten über das liefern, was zu dokumentieren ist.[75] Die Dokumentation unterscheidet sich von der *Transparenz*, da sie mit *interner Kommunikation* einhergehen kann, z. B. zwischen Mitarbeiterinnen und Mitarbeitern innerhalb eines Data-Science-Teams oder eines gesamten Unternehmens,[76] ohne jedoch Benutzerinnen und Benutzern oder der breiten Öffentlichkeit ausreichende Transparenz zu bieten.

Zweitens beinhaltet die Kontrolle das *Messen*, *Erfassen*, *Bewerten*[77] und *Definieren von Standards*[78] *und Richtlinien* für alle KI-bezogenen Prozesse und Ergebnisse. Sie umfasst daher, zu *untersuchen*, was erforderlich ist, um *aussagekräftige* Standards und Maßnahmen zu erzeugen, d. h. solche, die zu einem authentischen *Verständnis* und *Wissen* über die zu bewertenden Prozesse und Ergebnisse führen, die sowohl Produkte als auch Aspekte der Kontrolle über KI sind. KI-Systeme zu verstehen und zu kennen, erfordert, die Funktionsweise der Systeme zu *erklären*.[79] Daher sind die Ziele der *erklärbaren*[80] *KI* und der Wert der Erklärbarkeit in den Richtlinien der EU-Expertengruppe eine Facette (möglicherweise die wichtigste Facette) des Erfordernisses der prozessualen Kontrolle. KI-Systeme zu bewerten und einzuschätzen, erfordert nicht nur, Entwicklungsentscheidungen zu *rechtfertigen,*[81] sondern dies auch bei Fehlern, Vorurteilen und Güterabwägungen mit anderen moralischen Zielen zu tun, wenn sie unvermeidbar sind.

Drittens umfasst die Kontrolle die sozialen Aktivitäten, die erforderlich sind, um sicherzustellen, dass die Untersuchung der Prozesse und Ergebnisse angemessen vollständig ist und relevante Perspektiven nicht ausgeschlossen werden. Dies beinhaltet Aktivitäten wie *Schulung*[82] und Verbesserung des

72 Dataethical Thinkdotank, 2021, Traceability.

73 Council of Europe, CM/Rec (2020)1, Testing.

74 Council of Europe, CM/Rec (2020)1, Interaction of systems.

75 Bundesrat Leitlinien KI 2020. Leitlinie 3; Engelmann/Puntschuh, 2020, 7. Wirkungsmonitoring.

76 Engelmann/Puntschuh, 2020, 8. Nachvollziehbarkeit.

77 AI Now Institute et al., Key Elements Of A Public Agency Algorithmic Impact Assessment, #1; Reisman et al., 2018; Council of Europe, CM/Rec (2020)1, Ongoing review, Evaluation of datasets and system externalities, Testing on personal data; WEF, 2020, Data Quality; New York City, 2019, Impact determination.

78 Council of Europe, CM/Rec (2020)1, Standards; Engelmann/Puntschuh, 2020, 1. Zielorientierung.

79 Mittelstadt/Russell/Wachter, 2019; New York City, 2019, Explanation; Engelmann/Puntschuh, 2020, 8. Nachvollziehbarkeit.

80 Dataethical Thinkdotank, 2021, Explainability; Floridi/Crowls, 2019; Government of New Zealand, 2020, Transparency.

81 Leslie, 2019, Transparency; Council of Europe, CM/Rec (2020)1, Testing; Loi/Ferrario/Viganò, 2020.

82 Council of Europe, CM/Rec (2020)1, Personnel management; Engelmann/Puntschuh, 2020, 9. Akzeptanz.

*internen Fachwissens*,[83] *Überprüfung durch Expertinnen und Experten*[84] und sogar *Vielfalt in der Belegschaft*[85] sowie Transparenz als öffentliche Debatte,[86] wenn sie dazu dienen soll, das Verständnis einer Organisation für die sozialen Auswirkungen von KI-Systemen zu verbessern.

Viertens umfasst die Kontrolle Maßnahmen zur Risikominderung, beispielsweise, Backups und Notfallpläne zu erstellen,[87] Prozesse einzugrenzen und zu trennen, zu blockieren und zu unterbrechen,[88] die Möglichkeit für menschliches Eingreifen zu schaffen,[89] Risiken vorherzusagen und zu verhindern, schädliche oder riskante Praktiken zu verbieten,[90] Prozesse anzufechten[91] und Fehler zu korrigieren.[92] Die Bedeutung, die der Risikobewertung und dem Risikomanagement[93] in den hier analysierten Leitlinien beigemessen wird, kann kaum überschätzt werden.

Fünftens und mit besonderer Bedeutung für den Einsatz von KI-Systemen im *öffentlichen* Sektor beinhaltet die Kontrolle, wer über *Schlüsselinfrastrukturen*[94] Bescheid weiß, wem sie gehören und wer sie effektiv kontrolliert – z. B. die Datenbestände und Algorithmen für maschinelles Lernen, die wesentliche Voraussetzung dafür sind, aus Daten zu lernen und die eingesetzte KI weiterzuentwickeln, zu gestalten und zu kontrollieren.

Zu betonen ist, dass in der hier skizzierten Richtlinie kein Instrument zur Risikobewertung zur Verfügung gestellt wird, mit dem das Risiko *quantifiziert* werden soll. Die vorliegend untersuchten Werkzeuge, die von den Verwaltungen Kanadas[95] und Neuseelands[96] verwendet oder in den Richtlinien des Weltwirtschaftsforums[97] empfohlen werden, weisen allesamt Merkmale auf, die als problematisch erachtet werden können – insbesondere, wenn sie eine Rechtsgrundlage schaffen sollen, auf deren Basis bei Verstößen sanktioniert werden soll. Das neuseeländische Risiko-Tool stützt sich vollständig auf subjektive Risikobewertungen und fordert Benutzerinnen und Benutzer auf, anzugeben, ob ein Risiko «gelegentlich» besteht, «unwahrscheinlich» oder «wahrscheinlich» ist die Auswirkungen «gering», «mäßig» oder «hoch» sind. Jedoch werden keine konkreten, objektiven Kriterien angegeben, auf die sich diese Einschätzungen stützen können. Bezeichnungen wie «nicht ernst», «mäßig», «weit verbreitet» oder «ernst» sind sehr kontextbezogen und vage. In einem Kontext, in dem die Einstufung einer Anwendung als «kann schwerwiegende Konsequenzen haben» bedeutet, dass kostspielige Dokumentations- und Verwaltungsanforderungen folgen, ist zu erwarten, dass die Unbestimmtheit der Sprache Benutzerinnen und Benutzer dazu veranlasst, die Beschreibung des Risikos herunterzuspielen (wenn sie die Konsequenzen tragen müssen) oder einheitlich die höchste Risikostufe auszuwählen (wenn die Personen, die für das Risiko verantwortlich

83 AI Now Institute et al, Executive Summary; Council of Europe, CM/Rec (2020)1, Independent research and Rights-promoting technology.

84 AI Now Institute et al., 2018, Key Elements Of A Public Agency Algorithmic Impact Assessment, #2; Government of Canada, 2019, Appendix C; Council of Europe, CM/Rec (2020)1, Consultation and adequate oversight and Expertise and oversight.

85 Council of Europe, CM/Rec (2020)1, Principle of Equality and Security and Personnel management.

86 Council of Europe, CM/Rec (2020)1, Public debate.

87 Government of Canada, 2019, Appendix C.

88 Council of Europe, CM/Rec (2020)1, Follow up, Consultation and adequate oversight.

89 Government of New Zealand, 2020; Council of Europe, 2020, Ethical Charter AI, Principle under user control.

90 Council of Europe, CM/Rec (2020)1, Consultation and adequate oversight and Follow up.

91 Council of Europe, CM/Rec (2020)1, Barriers and Effective remedies.

92 Council of Europe, CM/Rec (2020)1, Consultation and adequate oversight and Effective remedies.

93 WEF, 2020, Key variables to consider in a risk assessment; Council of Europe, CM/Rec (2020)1, Human Rights Impact Assessment; Government of New Zealand, 2020, Assessing likelihood and impact, Human oversight and accountability, Reliability, Security and Privacy; Government of Canada, 2019, Algorithmic Impact Assessment.

94 Council of Europe, CM/Rec (2020)1, Infrastructure and Interaction of systems.

95 Government of Canada, 2019, Appendix B.

96 Government of New Zealand, 2020, Assessing likelihood and impact.

97 WEF, 2020.

sind, nicht die höheren Kosten für dessen Management tragen müssen). Darüber hinaus berücksichtigt das Auswirkungskriterium sowohl die Ausbreitung (Anzahl der Betroffenen) als auch die Schwere (wie schwerwiegend der Schaden ist) und lässt ungeklärt, wie diese Dimensionen zum Risiko beitragen.

Aus ähnlichen Gründen sind die Folgenabschätzungsniveaus der kanadischen «Richtlinie über automatisierte Entscheidungsfindung»[98] problematisch, da sie keine Kriterien zur Unterscheidung zwischen «geringer», «mäßiger», «hoher» und «sehr hoher» Auswirkung bietet; außerdem kann das Wort «oft» im Ausdruck «wird oft dazu führen» zu einer Vielzahl widersprüchlicher Interpretationen führen.

In den Richtlinien des Weltwirtschaftsforums heißt es: «Es ist wichtig, Faktoren wie die Anzahl der Betroffenen zu berücksichtigen.» Unklar ist jedoch, ob dies mit den anti-utilitaristischen Grundsätzen vereinbar ist, die die Gesetze vieler EU-Länder beeinflussen. Diese Art von Risikoindikator könnte darauf hindeuten, dass ein geringes Maß an Kontrolle über einen riskanten Algorithmus, der einem Individuum erheblichen Schaden zufügen kann, ethisch zulässig ist, solange nur die Würde einiger *weniger* Individuen von ihnen negativ beeinflusst wird (während die Mehrheit von ihrer Einführung profitiert). Dies steht auch im Spannungsfeld mit einer der Empfehlungen des Europarates, die bezogen auf die «Verantwortung der Akteure des Privatsektors in Bezug auf Menschenrechte und Grundfreiheiten im Kontext algorithmischer Systeme» eine klare *anti-utilitaristische* ethische Sichtweise annimmt, indem sie klar sagt («1.2. Umfang der Maßnahmen»), dass die Verantwortung für die Achtung der Menschenrechte «unabhängig von ihrer Größe» gilt (auch wenn Umfang und Komplexität der Mittel unterschiedlich sein können).[99]

Bei der Entwicklung der Leitlinien für den Kanton Zürich wurde ebenfalls davon ausgegangen, dass Kontrolle einen hohen Stellenwert besitzt. Diese wird vor allem durch *Dokumentation* umgesetzt. Einen Transparenzbericht zu schreiben, soll nicht nur oder nicht einmal hauptsächlich die externe Kontrolle und Überprüfung ermöglichen, sondern auch und vor allem die Verwaltung verpflichten, klar strukturierte und zielgerichtete Dokumentations-, Mess- und Bewertungsaufgaben zu übernehmen. *Frage 2.1* und *2.2* erfordern die Dokumentation der Ziele des Systems, *Frage 2.13* und *Frage 2.15* verlangen die Dokumentation der *Verantwortlichkeiten* und der *menschlichen Kontrollstruktur*, die zur Steuerung des Systems zur Verfügung stehen. *Fragen 2.7* und *2.8* und *Fragen 2.10, 2.12, 2.16, 2.17, 2.18* und *2.19* erfordern die Dokumentation von *Definitionen, Standards, Tests, Messungen und Bewertungen* der Leistung, des Datenschutzes, der Fairness und der beteiligten Stakeholder. *Fragen 2.9, 2.10, 2.14, 2.15* und *2.20* setzen die Dokumentation von *Risikomanagementstrukturen* einschließlich *Überwachung, Feedback, Fehlerkorrektur und Cybersicherheit* sowie ihrer Ergebnisse voraus.

Die Checkliste erfordert daher nicht, dass die Verwaltung den *Grad* des Risikos *quantifiziert*, bevor sie sich damit befasst. Sie funktioniert jedoch trotzdem als eine Art Risikobewertungsinstrument, da sie bestimmte Dokumentationsaufgaben unter der Bedingung notwendig macht, dass entsprechende Risikosignale von der Person, die die Checkliste beantwortet, erkannt werden. Je höher die Anzahl der Risikosignale ist, desto länger und strukturierter sind die Dokumentations- und Transparenzanforderungen und der Aufwand für die Verwaltung bei der Kontrolle der KI-Systeme. Sie ist eher als Checkliste für die *Risikoreaktion* als für die *Risikobewertung* gedacht.

## 2. Transparenz

Unter Transparenz werden hier ausschließlich die Vorlage und die Übermittlung von Informationen an Parteien ausserhalb der Institution verstanden, die eine KI-Lösung entwerfen oder implementieren, einschließlich der Wirtschaftsprüfer, externen Expertinnen und Experten, Journalistinnen und Journalisten, Politikerinnen und Politiker, Verantwortlichen in anderen Bereichen der Verwaltung und der breiten

98 Government of Canada, 2019.

99 Council of Europe, CM/Rec (2020)1, Appendix B.

Öffentlichkeit. Die Kommunikation zwischen den Mitgliedern des Data-Science-Teams, zwischen einem Data-Science-Team und der/dem CEO des Datenanalyseunternehmens, dem sie Bericht erstatten, oder zwischen einem privaten Unternehmen und der für die Beschaffung zuständigen Verwaltungseinheit wird hier als *intern* verstanden, da davon ausgegangen werden kann, dass alle diese Parteien in Bezug auf die Verwendung und Anwendung von KI-Systemen das gleiche Ziel verfolgen.

Es gibt mindestens vier Haupttheorien, warum Transparenz aus ethischer Sicht instrumentell wertvoll ist.[100] Erstens wird die Ansicht vertreten, dass «Sonnenlicht das beste Desinfektionsmittel ist», um Louis Brandeis zu zitieren, das bedeutet die Auffassung, dass Transparenz die Rechenschaftspflicht fördert, was wiederum verhindert, dass zumindest das schlimmste unethische Verhalten auftritt. Zweitens besteht die Meinung, dass Transparenz zur Qualität der Technologie beiträgt, da sie das Crowdsourcing von Expertenmeinungen und das Feedback der betroffenen Bürger ermöglicht, was zu einer besseren Prüfung der Technologie führt und sie vertrauenswürdiger macht. Drittens gibt es die Ansicht, dass Transparenz Endbenutzern einer Technologie oder Personen, die davon betroffen sein könnten, ermöglicht, eine fundierte Entscheidung darüber zu treffen, ob sie verwendet werden soll. Viertens wird vertreten, dass Transparenz eine öffentliche Debatte ermöglicht, die für die demokratische Legitimität technologischer Lösungen erforderlich ist, was besondere Relevanz hat, wenn die Implementierung von Technologie nicht wertneutral ist. Alle diese vier Rollen der Transparenz spiegeln sich in den untersuchten Richtlinien wider und es wird allgemein angegeben, dass sie zum Vertrauen in die Technologie beitragen.

Die Plausibilität, Stärke und Reichweite jeder dieser Theorien sind jedoch umstritten.[101] Die Rechenschaftstheorie ist möglicherweise nicht in allen Kontexten, sondern nur in solchen gültig, die aus dem einen oder anderen Grund die öffentliche Aufmerksamkeit auf sich ziehen. Die Crowdsourcingtheorie unterstützt möglicherweise nur schwächere Formen der Transparenz (die Technologie wird der Kontrolle einer begrenzten und ausgewählten Gruppe von Expertinnen und Experten unterworfen). Die Autonomietheorie kann die begrenzten kognitiven Ressourcen des Individuums, das sich für eine Technologie entscheiden muss, nicht angemessen berücksichtigen und ist nicht anwendbar, wenn Individuen nicht die Wahl haben, von der Technologie betroffen zu sein. Die Theorie der öffentlichen Debatte kann ihren Zweck verfehlen, wenn die breite Öffentlichkeit entweder nicht interessiert oder nicht qualifiziert genug

| Rolle der Transparenz | Ziele | Ethischer Wert |
|---|---|---|
| **Transparenz als Desinfektionsmittel** | Rechenschaftspflicht, Vermeidung von unethischem Verhalten | Schadensvermeidung |
| **Transparenz für Crowdsourcing** | Sammeln von Experten- und Laienmeinungen, Verbessern der Technologie | Nutzen |
| **Transparenz für fundierte Auswahl** | Informierte individuelle Auswahl ermöglichen | Autonomie (individuell) |
| **Transparenz für eine informierte öffentliche Debatte** | Informierte demokratische Deliberation ermöglichen | Autonomie (kollektiv), Demokratie |

Tab. 1 (eigene Darstellung): Transparenz

100 de Laat, 2017; Felzmann et al. 2019; Loi/Ferrario/Viganò, 2020; Zarsky, 2013.

101 Felzmann et al. 2019; Zarsky, 2013.

ist, um diese Art von Diskussion zu führen. Dennoch ist Transparenz das am häufigsten zitierte Prinzip in KI-Ethikrichtlinien: Nur sehr wenige Texte erwähnen sie nicht.[102] Unter den Richtlinien, die speziell für diesen Bericht überprüft wurden, gab es keine einzige, in der sie nicht genannt wurde.

Transparenz über KI-Systeme – oder besser gesagt das soziotechnische System, von dem KI nur ein Teil ist – herzustellen, wird in Bezug auf verschiedene Elemente erwartet: die Existenz automatisierter Entscheidungssysteme,[103] einschließlich ihres Zwecks, ihrer Reichweite und ihrer tatsächlichen Verwendung (*Frage 2.1* und *2.2*),[104] die Definitionen von Kernkonzepten und Kernmaßnahmen (z. B. der automatisierten Entscheidung oder der KI,[105] der Fairness[106]) (*Fragen 2.8, 2.10, 2.12*), die diesbezügliche ethische oder Folgenabschätzung,[107] ihre Rechtfertigung[108] (*Fragen 2.18* und *2.19*), die zugrunde liegenden Datentypen und Verarbeitungsmethoden[109] sowie deren Gesamtqualität, die als Genauigkeit,[110] Effektivität, Effizienz[111] oder Fähigkeit zur Unterstützung der Verwaltung[112] charakterisiert wird (*Fragen 2.16* und *2.17*). Im Vergleich zu früheren Richtlinien, die in vorherigen Analysen untersucht wurden,[113] scheint es weniger wichtig zu sein, individuelle Erklärungen der Ursachen oder Gründe zu liefern, warum eine bestimmte Entscheidung von einer KI getroffen wurde[114] – etwas, das im gerade entstehenden Feld der [e]X[plainable] KI im Vordergrund steht.

Während anerkannt wird, dass ein gleiches Maß an Transparenz nicht immer für alle Systeme angemessen ist,[115] sollte dieses jedoch so groß sein, wie es nach einer Güterabwägung mit anderen Zielen möglich ist.[116] Das Zielpublikum der Kommunikation kann variieren und die beteiligten oder betroffenen Personen,[117] die breite Öffentlichkeit[118,119] oder unabhängige Expertinnen und Experten umfassen.[120] Das Format der Kommunikation kann sich auch abhängig vom Kontext unterscheiden, obwohl dies selten spezifiziert wird, z. B. von allgemeinverständlichen Texten auf einer Webseite (z. B. die Erklärung, dass eine automatisierte Entscheidung getroffen wird) bis zu einem vollständig dokumentierten Bericht.[121] Mitunter wird sogar vorgeschlagen, Veröffentlichungen durch Whistleblower zu schützen und zu durch staatliche Gesetze und Strukturen in Unternehmen unterstützen.[122]

---

102 Jobin/Ienca/Vayena, 2019.

103 AI Now Institute et al.; Cities for Digital Rights, 2020; Council of Europe, CM/Rec (2020)1, Identifiability of algorithmic decision-making; Dataethical Thinkdotank, 2021, Fair communication.

104 AI Now Institute et al.; Engelmann/Puntschuh, 2020, 1. Zielorientierung; Engelmann/Puntschuh, 2020, 8. Nachvollziehbarkeit.

105 AI Now Institute et al.

106 Leslie, 2019.

107 AI Now Institute et al.; Council of Europe, CM/Rec (2020)1, Expertise and oversight; Government of Canada, 2019, Appendix C – Notice.

108 AI Now Institute et al.; Government UK, Transparency; Bundesrat Leitlinien KI, Leitlinie 3.

109 Council of Europe, 2020, Ethical Charter AI; Government of Canada, 2019, Appendix C – Notice.

110 Dataethical Thinkdotank, 2021, Fair communication.

111 Government of Canada, 2019, Reporting: 6.5.1.

112 Government of Canada, 2019, Appendix C – Notice.

113 Loi, 2020; Loi/Heitz/Christen, 2020.

114 Government UK, 2019, Transparency; Dataethical Thinkdotank, 2021, Transparency; WEF 2020, Human in the loop; Government of Canada, 2019, 6.2.3.

115 Council of Europe, CM/Rec (2020)1, Levels of transparency; Government of New Zealand, 2020, Transparency.

116 Council of Europe, CM/Rec (2020)1, Levels of transparency.

117 Dawson et al., 2020; Dataethical Thinkdotank, 2021, Transparency; Government UK, 2019, Ongoing review; Council of Europe, CM/Rec (2020)1, Expertise and oversight; Bundesrat Leitlinien KI 2020, Leitlinie 3.

118 Council of Europe, CM/Rec (2020)1, Public debate.

119 New York City, 2019, Available information.

120 Council of Europe, CM/Rec (2020)1, Expertise and oversight; AI Now Institute et al.; Bundesrat Leitlinien KI 2020, Leitlinie 3.

121 Government of Canada, 2019, 6.2, and Appendix C – Notice.

122 Council of Europe, CM/Rec (2020)1, Advancement of public benefit.

Transparenz kann als Eckpfeiler des hier skizzierten Ansatzes angesehen werden, bei dem Beamtinnen und Beamte einen Transparenzbericht erstellen müssen, der das wichtigste Ergebnis aller ethischen Aktivitäten darstellt. Nach der hier vertretenen Ansicht steht das gewünschte Maß an Transparenz in einem relativen Verhältnis zum Kontext und zur Art des zu beurteilenden KI-Systems. Anstatt eine Risikobewertung als Grundlage anzubieten, liefert der hier vorgeschlagene Ansatz Hinweise, mit welchen Aspekten von Transparenz die Verwaltung sich befassen muss bzw. welche Aspekte und Themen sie transparent machen muss.

## 3. Rechenschaftspflicht

Die Rechenschaftspflicht umfasst Maßnahmen, Entscheidungen, Rahmenbedingungen und Organisationsstrukturen, die die Verteilung und Identifizierung von Verantwortlichkeiten erleichtern sollen. Nur menschliche Personen können zur Rechenschaft gezogen werden, während dies bei einem KI-System nicht möglich ist. Rechenschaftspflichtige Akteure können Verantwortung für ihre Handlungen übernehmen und mit Sanktionen belegt werden. Die Förderung der Rechenschaftspflicht ist daher gleichbedeutend mit derjenigen der Fähigkeit, festzustellen, wer wofür verantwortlich ist und wer für unethische oder illegale Ergebnisse – nicht nur rechtlich, sondern auch organisatorisch oder durch Imageschäden – sanktioniert werden sollte.

Die im Rahmen dieser Studie untersuchten Richtlinien sehen vor, dass Rechenschaftspflichten wie folgt gefördert werden:

1. Indem Verantwortung zugewiesen wird – es sollte möglich sein, zu identifizieren, wer für die Gewährleistung ethischer Ergebnisse und Verhaltensweisen verantwortlich ist (vorausschauende Verantwortung)[123] und wer Sanktionen unterliegt, wenn dies nicht der Fall ist.[124]

2. Durch angemessene Strukturen und eine angemessene Organisation der Prozesse der Datenwissenschaft hinter dem KI-System und den Automatisierungsprozessen.[125] Organisationen sollten «eine kontinuierliche Verantwortungskette für alle Rollen einrichten, die am Entwurfs- und Lebenszyklus des Projekts beteiligt sind».[126] Hier ist zu beachten, dass eine kontinuierliche Verantwortungskette für alle Rollen erfordert, Prozesse und Ergebnisse klar zu dokumentieren, zu überwachen und zu kontrollieren.[127] Mit anderen Worten: Angemessene Strukturen für die Rechenschaftspflicht sind Kontrollstrukturen wie oben definiert (2.1). Der Kürze halber werden nicht alle Kontrollelemente wiederholt, die bereits in Ziffer 2.1 beschrieben wurden.

3. Indem ermöglicht wird, von scheinbar unpersönlichen Systemen getroffene Entscheidungen anzufechten oder abzulehnen[128] (in einigen Fällen kann das Rechtsstaatsprinzip herangezogen werden)[129] oder das KI-System grundsätzlich aufgrund seiner schädlichen oder diskriminierenden Auswirkungen anzufechten (mit Konzepten, die öffentlichen Beteiligungsverfahren entsprechen).[130]

4. Indem verlangt wird, dass Institutionen, die ein KI-System einsetzen, dafür verantwortlich sind, das Feedback der von ihm betroffenen

123 Dawson et al., 2020, Accountability; Bundesrat Leitlinien KI 2020, Leitlinie 4; Engelmann/Puntschuh, 2020, 4. Projektmanagement.
124 Government of New Zealand, 2020, Human oversight and accountability.
125 Engelmann/Puntschuh, 2020, 4. Projektmanagement.
126 Government UK, 2019, Accountability.
127 Government UK, 2019, Accountability.
128 AI Now Institute et al.; Council of Europe, CM/Rec (2020)1, Contestability; Dawson et al., 2020, Contestability.
129 AI Now Institute et al.
130 Cities for Digital Rights, 2020.

Personen zu sammeln und die erforderlichen Abhilfemaßnahmen umzusetzen.[131]

5. Indem verlangt wird, dass Schäden kompensiert werden, die aufgrund unethischen Verhaltens entstanden sind.

Die hier entwickelten praktischen Richtlinien greifen den Grundsatz der Rechenschaftspflicht auf, indem sie die Rechenschaftspflicht in ein Objekt der Transparenz verwandeln. Die *Fragen 2.3* bis *2.6* erfordern, dass die oben genannten individuellen Verantwortlichkeiten im Transparenzbericht angegeben werden. Die *Frage 2.15* und die *Frage 2.10* befassen sich mit der Existenz von Strukturen, mit denen Entscheidungen des KI-Systems infrage gestellt werden können.

131 AI Now Institute et al., Participatory Democracy, diversity and inclusion; Council of Europe, CM/Rec (2020)1, Consultation and adequate oversight; Dawson et al., 2020, Recourse; New York City, 2019, impact address.

# D. Checklisten

## I. Einleitung

Bei den folgenden zwei Checklisten handelt es sich um Hilfsmittel zur Herstellung von Transparenz bei technologischen Automationsprojekten und -anwendungen in der öffentlichen Verwaltung.

Die Methode zur Herstellung von Transparenz besteht darin, einen Transparenzbericht zu verfassen, der zeigt, dass die wichtigsten ethischen Fragen sowohl erkannt als auch unter menschliche Kontrolle gebracht wurden und eine angemessene Rechenschaftspflicht für den Prozess sichergestellt wurde.

Bei der Prüfung der ethischen Anforderungen für den Einsatz von KI-Systemen sollten zwei verschiedene Einschätzungen die praktischen Aktivitäten des Kantons Zürich leiten. Zu diesem Zweck werden zwei Checklisten zur Verfügung gestellt.

Die Checklisten greifen teilweise Aspekte auf, die bereits durch andere Instrumente abgedeckt sind. Außerdem können sich die Checklisten mit bestehenden Prozessen in der Verwaltung des Kantons Zürich überschneiden (zu denken ist etwa an Regeln und Abläufe zur Sicherstellung des Datenschutzes oder der Cyber-Security). Dies ist bei der Implementierung der Checklisten zu berücksichtigen, würde aber den Rahmen des vorliegenden Vorprojekts sprengen. Dasselbe gilt für die Koordination mit dem Bund.

Im Rahmen der Entwicklung dieser Checklisten wurde der Vorschlag geäußert, diese im Rahmen von HERMES (Initialisierungsphase) zu integrieren. Allerdings sind die Checklisten nicht nur bei der Durchführung eines Projekts, sondern auch bei späteren Änderungen zu berücksichtigen sind. Sie sind somit ab einem bestimmten Zeitpunkt nicht mehr für die Projektleitung, sondern für die Stammverwaltung maßgebend.

Anhand der ersten Checkliste (Triage-Checkliste) beurteilt die Verwaltung, welche ethischen Transparenzfragen während der Projektdurchführung im Detail zu dokumentieren sind, und wählt angemessene Vorgehensweisen für die Generierung derjenigen Daten und Bewertungen, die notwendig sind, um den Bericht mit informativem Inhalt zu füllen. Die folgenden Fragen helfen bei der Beurteilung:

— Mit wie vielen ethischen Transparenzaspekten muss die Verwaltung sich befassen?

— Wie viele ethische Transparenzverfahren müssen implementiert werden?

— Wie viele Ressourcen müssen für ethische Transparenzverfahren bereitgestellt werden?

— Welche Aspekte der ethischen Transparenz müssen im Bericht detailliert behandelt werden? (Und ist ein solcher Bericht überhaupt notwendig?)

Die zweite Checkliste (Checkliste Transparenzbericht) dient als Leitfaden für die Erstellung eines ausführlichen Transparenzberichts (im Folgenden: Transparenzbericht).

Der Transparenzbericht kann erst am Ende einer Entwicklung und Implementierung eines KI-Systems (im Folgenden: Projekt) erstellt werden (einschließlich der Interaktion des soziotechnischen Systems mit der betroffenen Öffentlichkeit in Fällen, in denen die Beurteilung ethischer Fragen eine solche Überwachung erfordert). Allerdings muss mit der Erstellung

des Transparenzberichts bereits während des Projekts begonnen werden: Manche Informationen, die für den Transparenzbericht notwendig sind, können nur in den verschiedenen Phasen der Projektdurchführung und nicht erst nach Projektabschluss generiert werden. Die Checkliste Transparenzbericht enthält deshalb auch Hinweise darauf, in welcher Phase des Prozesses spezifische, für die Transparenz notwendige Informationen generiert werden müssen. Am Ende des Projekts muss der Transparenzbericht klare Informationen über die umgesetzten Prozesse enthalten, die zur Adressierung der in Checkliste 1 (Triage-Checkliste) hervorgehobenen spezifischen ethischen Punkte geeignet sind.

Falls sich die Prozesse nach der Einführung des Systems ändern, muss die Verwaltung überprüfen, ob die ursprüngliche Einschätzung noch gültig ist oder sich zusätzliche ethische Transparenzprobleme ergeben haben.

Diese ethischen Transparenzprozesse sollten von der Verwaltung eingeleitet werden und alle potenziellen Transparenzaspekte berücksichtigen, die in der Checkliste angegeben sind. Wenn die Verwaltung nicht in der Lage ist, ein angemessenes Maß an Transparenz über diese ethischen Fragen zu schaffen, oder wenn die Transparenz die Unzulänglichkeit des soziotechnischen Systems in einer Weise aufzeigt, die mit dem Ruf der Behörde, die das Projekt verantwortet, unvereinbar ist, sollte dies dazu führen, das Projektziel zu überdenken und/oder mehr Ressourcen in die Suche nach einer praktikablen Lösung zu investieren.

Unter Transparenz wird im vorliegenden Kontext die Kommunikation an verschiedene Zielgruppen verstanden. Im Falle einiger Anwendungen oder für einige Informationskategorien wird die breite Öffentlichkeit (jede/jeder) die Adressatin sein. Dies kann z. B. durch die Veröffentlichung eines Berichts mit allen Informationen auf einer Webseite umgesetzt werden. Bei anderen Anwendungen oder für einige andere Informationskategorien werden eine vorgesetzte Stelle/ein Verantwortlicher innerhalb derselben Abteilung oder einer anderen Abteilung, technische Expertinnen/Experten oder Vertreterinnen/Vertreter von Interessengruppen, denen die Dokumentation vertraulich mitgeteilt wird, die Zielgruppe sein.

# II. Checkliste 1: Triage-Checkliste für KI-Systeme

## 1. Einleitende Bemerkungen

Die Fragen in dieser Checkliste sollten so früh wie möglich, d. h. bereits in der Planungsphase, beantwortet werden da sie dabei helfen, neben den primären Projektzielen zusätzliche Spezifikationen für das Projekt zu berücksichtigen. Die Checkliste 1 hilft bei der Ermittlung der wichtigsten ethischen Transparenzaspekte, die dokumentiert werden müssen. Sie sollte nicht als erschöpfende Liste aller Risiken für alle Zusammenhänge angesehen werden (es können andere Risiken für Personen, Sachen oder die Gesellschaft insgesamt bestehen). Es handelt sich um eine methodische Herangehensweise, jedoch besteht keine Garantie.

Die erste Checkliste ist bewusst so formuliert, dass sie nicht nur für spezifische KI-Technologien gilt. Vielmehr soll mit ihrer Hilfe ermittelt werden, welche Arten von KI-Systemen einen (erhöhten) Transparenzbedarf aufweisen. Es ist einfacher, anhand der Checkliste ein adäquates Maß an Transparenz zu erreichen, wenn

— das Projekt ethische Aspekte während der frühen Planung und bei der Implementierung des Produkts berücksichtigt,

— die Spezifikationen zur Befassung mit solchen ethischen Aspekten von Anfang an einbezogen werden (*Ethics-by-Design*-Ansatz),

— die Informationen zu den zur Behandlung der ethischen Aspekte ergriffenen Maßnahmen in jeder Phase des Projekts ordnungsgemäß dokumentiert werden,

— die Transparenz und die Rechenschaftspflicht der von Menschen durchgeführten Verfahren in allen Phasen des Projekts bestehen.

Die Checkliste impliziert zwei Ebenen der Transparenz: Transparenz mit geringem Detaillierungsgrad und Transparenz mit hohem Detaillierungsgrad:

— Niedriger Detaillierungsgrad der Transparenz: Dokumentieren und speichern Sie nur die Antworten auf die Triage-Checkliste einschließlich der Begründungen für Ihre Antworten.

— Hoher Detaillierungsgrad der Transparenz: Legen Sie einen detaillierten Transparenzbericht ab (Checkliste 2). Die Antworten auf die Fragen der Triage-Checkliste bestimmen, ob ein Transparenzbericht ausgefüllt werden muss und welche Abschnitte des Transparenzberichts ausgefüllt werden müssen.

Bewertungsstufe für die Triage-Checkliste: ganz am Anfang eines Projekts.

## 2. Schadensvermeidung

1.1. Befasst sich die Entscheidung mit speziellen Kategorien personenbezogener Daten?

1.2. Haben böswillige Parteien besonders starke Motive, das System zu hacken? Können sie, auch durch Erpressung, einen einfachen und substanziellen finanziellen Gewinn erzielen oder kann ein gehacktes System verwendet werden, um politische Ziele zu erreichen (einschließlich der Äußerung politischer Opposition gegen das System)?

1.3. Wird das soziotechnische System verwendet, um Entscheidungen über Personen zu treffen, zu empfehlen oder zu beeinflussen, und zwar in einer Weise, die Auswirkungen darauf hat, welche Entscheidung getroffen wird?

[Z. B. ist eine automatische Rechtschreibprüfung Teil eines soziotechnischen Systems. Wenn es von Menschen verwendet wird, die Entscheidungen über Individuen treffen, kann es als «verwendet, um Entscheidungen über Individuen zu treffen» beschrieben werden. Aber es beeinflusst nicht (in irgendeiner erkennbaren und wissenschaftlich plausiblen Weise), welche Entscheidung getroffen wird.]

1.4. Wird das System verwendet, um eine Entscheidung über eine gesetzliche Pflicht oder ein Recht einer Person zu treffen? [Zur Bedeutung von «für eine Entscheidung verwendet» in diesem Zusammenhang siehe Anleitungstext zu *Frage 1.3*.]

1.5. Macht es das System mehr oder weniger wahrscheinlich, dass bestimmte Personen den Wesensgehalt eines Rechts genießen? Oder macht es das System mehr oder weniger wahrscheinlich, dass bestimmte Personen sanktioniert werden? Oder beeinflusst das System die Wahrscheinlichkeit, dass ein Einzelfall die Aufmerksamkeit der Verwaltung auf sich zieht oder von dieser ignoriert wird?

1.6. Können Einzelpersonen die Entscheidung vermeiden oder verlangen, dass die Entscheidung über ein anderes Verfahren getroffen wird, bei dem nicht dasselbe technische System verwendet wird?

1.7. Kann die Person, über die mithilfe des Tools eine Entscheidung getroffen wurde, beweisen, dass diese falsche ist, ohne vor Gericht zu gehen?

1.8. Ist der Schaden einer falschen Entscheidung vollständig reversibel?

1.9. Ist es möglich, den Einzelnen oder die Familie vollständig und angemessen zu entschädigen, wenn festgestellt wird, dass die Entscheidung falsch war und irreversibel ist?

1.10. Betrifft die Entscheidung einen der folgenden Bereiche des öffentlichen Lebens oder Ressourcen des öffentlichen Sektors:

— die Rechtspflege,

— den Zugang zu Bildungschancen,

— den Zugang zu demokratischen Prozessen,

— den Zugang zur Gesundheitsversorgung,

— Maßnahmen im Bereich der öffentlichen Gesundheit,

— die Umwelt?

— 1.11. Findet durch die Anschaffung bzw. den Einsatz des KI-Systems in einem der

folgenden Bereiche eine Änderung statt bei:

— öffentlicher Computer-Infrastruktur,

— öffentlichen Datenbeständen oder

— immateriellen Vermögenswerten (z. B. Kompetenzen) im öffentlichen Sektor?

### 3. Gerechtigkeit und Fairness

1.12. Besteht das Risiko, dass das System eine politische Entscheidung (z. B. Wahl oder Volksabstimmung) beeinflusst?

1.13. Beeinflusst das technische System die Verteilung öffentlicher Mittel an wirtschaftliche Akteure in der Gesellschaft?

1.14. Beruht das technische System auf einem statistischen Modell des menschlichen Verhaltens oder der persönlichen Merkmale?

1.15. Ist das System so konzipiert, dass es adaptiv ist, damit nicht alle neuen Fälle wie andere behandelt werden, denen es in der Vergangenheit begegnet ist, weil es seine Parameter ändert, z. B. um effizienter zu werden?

### 4. Autonomie

1.16. Ist es das Ziel des technischen Systems, ein vollständig deterministisches Regelsystem zu automatisieren, das nur ein Minimum an Kreativität und menschlichem Urteilsvermögen durch die derzeitigen menschlichen Anwenderinnen/Anwender erfordert und keine Risiko- oder Wahrscheinlichkeitsabschätzungen beinhaltet?

1.17. Beruht das technische System auf Parametern, Merkmalen, Faktoren oder Entscheidungskriterien, die nicht den von den meisten Fachleuten auf diesem Gebiet normalerweise berücksichtigten Aspekten entsprechen?

1.18. Beurteilt das technische System (durch Vorhersagen oder Empfehlungen), dass den zuständigen Mitarbeitern der öffentlichen Verwaltung die Kompetenz (im Gegensatz zur Befugnis) zur Kritik und zum Verwerfen einer Entscheidung fehlt?

1.19. Greift das technische System auf die Infrastruktur eines Drittanbieters zurück, über die die öffentliche Einrichtung keine uneingeschränkte Kontrolle und/oder bei der sie keinen Zugriff auf z. B. Datensätze oder die Rechenleistung hat?

## III. Checkliste 2: Transparenzbericht

Das Ziel der zweiten Checkliste ist, Transparenz herzustellen und damit die Vertrauenswürdigkeit des Prozesses zu fördern. Das Ziel wird mittels Erstellung eines Berichts erreicht.

**Benutzungsanweisung:** Wenn die Beantwortung der Checkliste 1 ergibt, dass ein Transparenzbericht zu verfassen ist, wird mithilfe des Flussdiagramms bestimmt, welche Fragen aus Checkliste 2 der Transparenzbericht beantworten muss.

### 1. Abschnitt: Bewertungsphase für die Fragen 2.1. bis 2.6.: Bevor Sie Ihr System entwerfen

**Transparenz hinsichtlich von Werten**

2.1. Für welches Problem soll das System eine Lösung liefern?

2.2. Welches sind die weiteren Anforderungen des Systems? Berücksichtigen Sie zumindest:

— 2.2.1. Privatsphäre
[Bitte gehen Sie in diesem Teil des Berichts spezifisch auf die in Checkliste 1 – *Frage 1.1.* genannten Aspekte ein]

— 2.2.2. Cybersicherheit
[Bitte gehen Sie in diesem Teil des Berichts spezifisch auf die in Checkliste 1 – *Frage 1.2.* genannten Aspekte ein]

— 2.2.3. Fairness
[Bitte gehen Sie in diesem Teil des Berichts spezifisch auf die in Checkliste 1 – *Fragen 1.12. bis 1.15.* genannten Aspekte ein]

— 2.2.4. Erklärbarkeit
[Bitte gehen Sie in diesem Teil des Berichts spezifisch auf die in Checkliste 1 – *Fragen 1.17. und 1.18.* genannten Aspekte ein]

**Transparenz der Rechenschaftspflicht**

[Es wird empfohlen, diesen Abschnitt in jedem Fall auszufüllen. Wenn *Frage 1.16.* mit «Ja» beantwortet wurde, muss aufgezeigt werden, dass die neuen Zuständigkeiten mindestens gleichwertige Möglichkeiten zur Anfechtung von Entscheidungen bieten wie das zuvor bestehende System.]

2.3. Wer ist für die Konstruktion des Systems verantwortlich (Ebene Projektorganisation)?

2.4. Wer ist für den Einsatz des Systems und dessen Resultate verantwortlich (Ebene Stammorganisation)?

2.5. Wer ist verantwortlich für die Verwaltung der Antworten und Rückmeldungen der Endbenutzerinnen/Endbenutzer, d. h. der Personen, die das System benutzen oder von ihm unterstützt werden?

2.6. Wer ist dafür verantwortlich, auf Zweifel oder Herausforderungen des Einzelnen, der von der Nutzung des Systems betroffen ist, zu antworten?

## 2. Abschnitt: Bewertungsphase für die Fragen 2.7. bis 2.19: Nach dem Testen des Systems

**Transparenz der Umsetzung und der Steuerung**

2.7. Mit welchen Methoden wurde die Leistung des Systems getestet und gemessen?

[Bitte geben Sie an, wie Sie die Leistung in Bezug auf das in Checkliste 2 – *Frage 2.1.* angegebene Hauptziel messen.]

2.8. Welche Methoden wurden verwendet?

2.8.1. Welche Methoden wurden verwendet, um die von den Systemvorhersagen/-empfehlungen/-entscheidungen unmittelbar betroffenen Stakeholder zu identifizieren? Und was sind die voraussichtlichen Auswirkungen auf diese Personen?

2.8.2. Welche Methoden wurden verwendet, um die von der digitalen Transformation in der öffentlichen Verwaltung betroffenen Personen zu identifizieren (z. B. Personal der öffentlichen Verwaltung)? Und was sind die voraussichtlichen Auswirkungen auf diese Personen?

2.9. Welche Protokolle sind vorhanden, um Systemfehler und Fehlfunktionen zu behandeln?

2.10. Welche Methoden wurden zur Definition und zum Schutz der Privatsphäre verwendet?

[Bitte gehen Sie in diesem Teil des Berichts spezifisch auf die in Checkliste 1 – *Frage 1.1.* genannten Aspekte ein.]

2.11. Welche Maßnahmen zum Schutz der Cybersicherheit wurden getroffen?

[Bitte gehen Sie in diesem Teil des Berichts spezifisch auf die in Checkliste 1 – *Frage 1.2.* genannten Aspekte ein.]

2.12. Welche Methoden wurden verwendet, um die Voreingenommenheit und die Fairness des Systems zu definieren und zu messen?

[Bitte gehen Sie in diesem Teil des Berichts spezifisch auf die in Checkliste 1 – *Fragen 1.12. bis 1.14.* genannten Aspekte ein, die in diesem Zusammenhang relevant sind.]

2.13. Wie werden den Systemendbenutzern und den Personen, die vom Einsatz des Systems unmittelbar betroffen sind, individuelle Vorhersagen/Empfehlungen/Entscheidungen des Systems erklärt?

[Bitte gehen Sie in diesem Teil des Berichts spezifisch auf die in Checkliste 1 – *Fragen 1.17.* und *1.18.* genannten Aspekte.]

2.14. Wird die Systembereitstellung nach der Testphase kontinuierlich überwacht?

a) Zu jedem Zeitpunkt?

b) In einem bestimmten Zeitpunkt?

c) Mit welchen Maßnahmen?

2.15. Gibt es Möglichkeiten für Personen, die von einer Entscheidung betroffen sind, den Output des automatisierten Systems zu erfahren und die vom System beeinflussten Vorhersagen/Empfehlungen/Entscheidungen anzufechten?

[Falls zutreffend, beschreiben Sie, wie diese Kanäle mit den organisatorischen Rollen zusammenhängen, die in Checkliste 2 – *Fragen 2.5.* und *2.6.* erwähnt werden.]

**Transparenz hinsichtlich von Leistungen**

Auf der Grundlage der bisherigen Testläufe:

2.16. Wie verhält sich das System in Bezug auf die ausgewählten relevanten Metriken?

[Bitte beachten Sie alle Ziele und Anforderungen, die in Checkliste 2 – *Fragen 2.1.* angegeben sind.]

2.17. Wie ist das System im Vergleich zu dem zuvor vorhandenen, falls zutreffend, oder mit etablierten Benchmarks, falls vorhanden?

[Bitte beachten Sie *auch* die Ziele und Anforderungen, die in Checkliste 2 – *Fragen 2.1.* und *2.2.* angegeben sind.]

2.18. Welches sind die verbleibenden Sicherheits- und Datenschutzrisiken und warum sind sie angemessen?

[Bitte gehen Sie in diesem Teil des Berichts auf alle Anforderungen ein, die in Checkliste 2 – *Fragen 2.1.* und *2.2.* sowie *2.10.* und *2.11.* genannt werden.]

2.19. Bitte beschreiben Sie relevante ungelöste Voreingenommenheit oder mögliche Ursachen für Ungerechtigkeiten im System und erklären Sie, warum sie nicht gelöst werden können (beispielsweise, indem Sie Kompromisse mit anderen Systemzielen einschließlich widersprüchlicher Fairnessziele erläutern).

[Bitte gehen Sie in diesem Teil des Berichts auf alle Anforderungen ein, die in Checkliste 2 – *Frage 2.3.* genannt werden, und erläutern Sie, wie die in Checkliste 1 – *Fragen 1.12. bis 1.15.* identifizierten Aspekte adressiert wurden.]

## 3. Abschnitt: Bewertungsphase für die Frage 2.20: Nach der Implementierung des Systems, wenn das System überwacht wird

Die Transparenz über die angebotenen Lösungen zu den ethischen Fragen erfordert manchmal eine kontinuierliche Überwachung des Projekts auch über die Testphase hinaus. Kontinuierliche Überwachung bedeutet, dass der Transparenzbericht aktualisiert werden muss.

2.20. Wurden während der Überwachung Vorhersagen/Empfehlungen/Entscheidungen des Systems jemals

a) vom Systemendbenutzer oder

b) von Personen, die Entscheidungen unterliegen, hinterfragt?

## IV. Beispiel für den Einsatz der Checklisten 1 und 2: *Swiss COMPAS*

Im Folgenden wird anhand eines imaginären Tools – *Swiss-COMPAS*-System zur Prognose von Rückfällen (ähnlich dem US-amerikanischen System *COMPAS*)[132] – die Anwendung der Triage für KI-Systeme (Checkliste 1) und der Vorgaben für den Transparenzbericht (Checkliste 2) veranschaulicht. Dieses imaginäre System wäre eine Anwendung, mit dem anhand von Daten über eine inhaftierte Person die Wahrscheinlichkeit bestimmt wird, dass diese nach ihrer Entlassung aus dem Gefängnis erneut straffällig wird.

### 1. Vorbemerkungen

Dieser hypothetische Bericht wird nur als Illustration dafür zur Verfügung gestellt, wie ein auf Tatsachen basierender Bericht aussehen würde. Der vorliegende Fall soll kein Modell für «ethische Best Practices» darstellen: Transparenz wird erreicht, wenn die Situation in Bezug auf die Automatisierung einschließlich der Aspekte, die möglicherweise nicht auf ethisch angemessene Weise behandelt worden sind, ehrlich dargestellt wird. Anzunehmen ist, dass die meisten Praxisfälle unter dem Gesichtspunkt der ethischen Angemessenheit unvollkommen sein werden: Transparenz soll schrittweise, sukzessive Verbesserungen ermöglichen.

Dieser hypothetische Transparenzbericht wird als imaginärer «erster Entwurf» der zuständigen Behörden verstanden, der durch den Austausch mit Stakeholdern, die mehr Informationen verlangen, als der aktuelle Bericht bereitstellt, verbessert wird. Transparenz soll den Einsatz von KI-Technologien nicht abschließend rechtfertigen, sondern diese schrittweise verbesserungsfähig machen.

Das Flussdiagramm unterstützt Nutzerinnen und Nutzer dabei, den Transparenzbericht in einer anderen Reihenfolge als derjenigen der verschiedenen Abschnitte im Abschlussbericht auszufüllen.

Die Struktur des Transparenzberichts ist immer dieselbe, aber die Triage-Fragen und die Flussdiagrammbefehle bestimmen, welche Abschnitte in einem bestimmten Fall ausgefüllt werden müssen und welche nicht. Dies kann von Fall zu Fall unterschiedlich sein und hängt davon ab, wie die Triage-Fragen beantwortet werden.

Das Flussdiagramm kann angeben, dass bestimmte Fragen mehr als einmal beantwortet werden müssen (z. B. als Antwort auf verschiedene Triage-Fragen). Solche Wiederholungen können ignoriert werden. In diesem Beispiel wird keine Wiederholung erwähnt, die durch die Anforderungen des Flussdiagramms erzeugt wird.

Der Abschnitt «Checklisten 1 und 2» stellt dar, wie eine Person oder Gruppe, die verantwortlich dafür ist, einen Transparenzbericht zu erstellen, die Fragen auf Grundlage der Checklisten beantworten könnte. Der Abschnitt «Transparenzbericht» stellt den Bericht in einer Fassung dar, die veröffentlicht werden könnte.

132 https://en.wikipedia.org/wiki/COMPAS_(software).

## 2. Checklisten 1 und 2

| *Frage aus der Triage-Checkliste (Checkliste 1)* | *Antwort auf die Frage* | *Konsequenz für den Transparenzbericht (ausgehend von den Fragen in Checkliste 2)* |
|---|---|---|
| **1.1. Befasst sich die Entscheidung mit speziellen Kategorien personenbezogener Daten im Sinne des kantonalen Rechts?** | Das (hypothetische) Schweizer *COMPAS*-Bewertungstool erfordert, das Geschlecht der bewerteten Person zu erheben. Diese Angabe ist durch das Diskriminierungsverbot geschützt bzw. in bestimmten Fällen untersagt. Es werden zudem vertrauliche Informationen gesammelt, z. B., ob die untersuchte Person eine Trennung ihrer Eltern durchlebt hat, oder Strafregistereinträge von Freunden und Verwandten. | 2.2.1. Welche Anforderungen werden an das System in Bezug auf die *Privatsphäre* gestellt?<br>[Legen Sie hier die Anforderungen an das System dar. Welche Maßnahmen bezüglich des Datenschutzes sollten ergriffen werden? Nehmen Sie beispielsweise Kontakt mit der/dem Datenschutzbeauftragten Ihrer Organisation auf, um diesen Teil des Berichts zu verfassen.]<br>2.8.1. Welche Methoden wurden verwendet, um die von den Systemvorhersagen/-empfehlungen/-entscheidungen unmittelbar betroffenen Stakeholder zu identifizieren? Und was sind die voraussichtlichen Auswirkungen auf diese Personen?<br>Beispiel: «Wir haben ein Brainstorming-Meeting mit Staatsanwälten und Richtern des Kantons und Anwälten der Strafjustiz durchgeführt. Bei diesem Treffen wurden die Stakeholder identifiziert, die direkt von den Vorhersagen betroffen sind, nämlich die Angeklagten, ihre Verteidiger, ihre Familien, potenzielle künftige Opfer, wenn die Angeklagten erneut straffällig werden, und Gemeinschaften, in denen Menschen, die möglicherweise erneut straffällig werden, leben.<br>Unsere Analyse der Stakeholderinteressen sieht wie folgt aus:<br>A) Angeklagte. Das System zu verwenden, liegt im Interesse der Angeklagten, bei denen es unwahrscheinlich ist, dass sie erneut straffällig werden (oder die statistisch nicht von denjenigen zu unterscheiden sind, bei denen es unwahrscheinlich ist, dass sie erneut straffällig werden). Es ist insbesondere im Interesse derjenigen, die ihr Recht auf ein günstiges Bewährungsurteil am schlechtesten ausüben können, da sie sich nicht die besten Anwälte leisten können. Es ist nicht im Interesse von Personen, die sich mithilfe guter Anwälte bessere Chancen verschaffen können, auf Bewährung entlassen zu werden.<br>B) Anwälte der Angeklagten. Das System ist nicht in ihrem Interesse, da es ein Bestandteil richterlicher Entscheidung sein wird, den die Anwälte nicht anfechten können. ▶ |

| *Frage aus der Triage-Checkliste (Checkliste 1)* | *Antwort auf die Frage* | *Konsequenz für den Transparenzbericht (ausgehend von den Fragen in Checkliste 2)* |
|---|---|---|
| | | C) Familien. Die Familien der Angeklagten werden dann von der höheren Wahrscheinlichkeit profitieren, dass der Angeklagte auf Bewährung freigelassen wird, wenn der Einsatz von Swiss *COMPAS* im Vergleich zum Status quo zu einem höheren Anteil an gewährten Bewährungsstrafen führt (es sei denn, Angeklagte sind wegen eines Verbrechens gegen ihre Familien angeklagt). Dies hängt eng mit der Verhältnismässigkeit der Entscheidungen zur Gewährung der Bewährung zusammen.<br>D) Potenzielle Opfer. Sie werden dann von Swiss *COMPAS* profitieren, wenn dadurch ein geringerer Anteil der wieder straffällig werdenden Angeklagten freigelassen wird. Dieses Interesse wird nicht unbedingt gefördert, wenn Swiss *COMPAS* zu einem geringeren Anteil an gewährten Bewährungsstrafen führt. Wenn das Tool einerseits weniger genau ist als menschliche Richter, kann ein geringerer Anteil der gewährten Bewährungsstrafen dazu führen, dass die Quote der Straftäter, deren Haftstrafe zur Bewährung ausgesetzt wurde, die aber erneut straffällig werden, steigt. Wenn das Tool andererseits jedoch genauer ist als menschliche Richter, kann ein höherer Anteil der gewährten Bewährungsstrafen damit einhergehen, dass weniger Verbrechen von Straftätern begangen werden, deren Haftstrafe zur Bewährung ausgesetzt wurde.<br>E) Das Interesse der Gemeinschaften, die davon profitieren könnten, kann als Kombination folgender Interessen angesehen werden:<br>der Familienmitglieder der Angeklagten, wie in Buchstabe C oben angegeben,<br>der Personen, deren Interessen mit denen der Familienmitglieder in Einklang stehen,<br>der Interessen der potenziellen Opfer von Straftätern, deren Haftstrafe zur Bewährung ausgesetzt wurde, wie oben in D angegeben,<br>der Personen, deren Interessen mit denen der potenziellen Opfer von Straftätern in Einklang stehen, deren Haftstrafe zur Bewährung ausgesetzt wurde (z. B. der Kinder des Opfers). ▶ |

| *Frage aus der Triage-Checkliste (Checkliste 1)* | *Antwort auf die Frage* | *Konsequenz für den Transparenzbericht (ausgehend von den Fragen in Checkliste 2)* |
|---|---|---|
| | | Ein Tool, das in der Lage ist, mehr Menschen auf Bewährung freizulassen, während gleichzeitig die Häufigkeit der erneuten Straffälligkeit von Straftätern, deren Haftstrafe zur Bewährung ausgesetzt wurde, abnimmt, sollte von den Gemeinschaften der Angeklagten begrüsst werden.»<br>2.10. Welche Methoden wurden zur Definition und zum Schutz der Privatsphäre verwendet?<br>[Hier legen Sie dar, welche Methoden tatsächlich in Bezug auf die Privatsphäre zum Schutz der personenbezogenen Daten eingesetzt worden sind.]<br>2.18. Welches sind die verbleibenden Sicherheits- und Datenschutzrisiken und warum sind sie angemessen?<br>[Hier erklären Sie, warum das Cybersicherheitsrisiko, das sich aus den Ziffern 2.9. und 2.11. ergibt, angesichts dessen, was auf dem Spiel steht, und der Wahrscheinlichkeit einer Attacke als angemessen beurteilt wird.] |
| **1.2. Haben böswillige Parteien besonders starke Motive, das System zu hacken? Können sie, auch durch Erpressung, einen einfachen und substanziellen finanziellen Gewinn erzielen oder kann ein gehacktes System verwendet werden, um politische Ziele zu erreichen (einschliesslich der Äusserung politischen Widerspruchs gegen das System)?** | Da die gesammelten Daten sensibel sind, muss die Cybersicherheit hoch sein, um Attacken durch motivierte Eindringlinge zu verhindern. | 2.2.2. Welche Anforderungen werden an das System in Bezug auf die *Cybersicherheit* gestellt?<br>[Legen Sie hier die Anforderungen an das System dar. Fordern Sie von Cybersicherheitsexpertinnen und -experten eine technische Expertise an.]<br>2.9. Welche Verfahren sind vorhanden, um Systemfehlern und Fehlfunktionen zu begegnen?<br>[Hier fügen Sie einen Abschnitt zur Cybersicherheit ein. Sie erläutern beispielsweise, wie Sie mit Fehlern von Mitarbeitern umgehen, die die Integrität, Verfügbarkeit oder Vertraulichkeit der gesammelten Informationen gefährden. Dieser Abschnitt wird am besten von Cybersicherheitsexperten entworfen.]<br>2.11. Welche Massnahmen zum Schutz der Cybersicherheit wurden getroffen?<br>[In diesem Abschnitt erläutern Sie die im System integrierten Cybersicherheitsmassnahmen. Dieser Abschnitt wird am besten von Cybersicherheitsexperten entworfen.]<br>2.18. Welche Methoden wurden verwendet, um die von den Systemvorhersagen/-empfehlungen/-entscheidungen unmittelbar betroffenen Stakeholder zu identifizieren? Und was sind die voraussichtlichen Auswirkungen auf diese Personen?<br>Antwort: siehe oben (Punkt 2.18 bei Frage 1.1.) |

| *Frage aus der Triage-Checkliste (Checkliste 1)* | *Antwort auf die Frage* | *Konsequenz für den Transparenzbericht (ausgehend von den Fragen in Checkliste 2)* |
|---|---|---|
| **1.3. Wird das sozio-technische System verwendet, um Entscheidungen über Personen zu treffen, zu empfehlen oder zu beeinflussen, und zwar in einer Weise, die beeinflusst, welche Entscheidung getroffen wird?** | Ja, das System liefert Ergebnisse, die von Richtern verwendet werden, um zu entscheiden, ob dem Angeklagten die Aussetzung der Haft zur Bewährung gewährt wird. | 2.2.3. Welche Anforderungen werden an das System in Bezug auf die *Fairness* gestellt?<br>[Dieser Aspekt ist zu komplex, um hier darauf einzugehen – die Beurteilung erfordert eine gemeinsame Analyse von zumindest Expertinnen und Experten für Statistik und Kriminologie, die in der Lage sind, eine begründete Einschätzung dazu zu geben, was sie für ein faires und unvoreingenommenes Urteil halten; im Idealfall sollten Strafverteidiger und/oder Expertinnen oder Experten für soziale Gerechtigkeit beauftragt werden oder die Ergebnisse überprüfen.]<br>Beispiel: Wir definieren Fairness wie folgt: Die Vorhersage, ob jemand erneut straffällig wird, ist im Durchschnitt für Männer und Frauen gleichermaßen genau.<br>Die Rechtfertigung dafür, Fairness so zu bestimmen, ist: …<br>2.2.4. Welche Anforderungen werden an das System in Bezug auf die *Erklärbarkeit* gestellt?<br>Beispielhafte Erwägungen: Da dies eine weitreichende Entscheidung ist, die letztendlich von Menschen (Richterinnen und Richtern) getroffen wird, erscheint es wichtig, dass sie ein mentales Modell der Faktoren und ihrer Gewichtung bilden können, die bei der Erzeugung des Scores berücksichtigt wurden. Sofern das Schweizer *COMPAS-Tool* einen geheimen Algorithmus verwendet, kann diese Anforderung möglicherweise nicht erfüllt werden. Wenn die Formel jedoch öffentlich bekannt wäre, würden um Bewährung ersuchende Personen unaufrichtig antworten, um ihre Risikobewertung zu verbessern. (Der US-*COMPAS*-Algorithmus ist geheim.)<br>Ein realer Bericht würde im Idealfall eine gründlichere Analyse und Diskussion dieses Widerspruchs sowie potenziell realisierbare Lösungen beinhalten.<br>2.7. Mit welchen Methoden wurde die Leistung des Systems getestet und gemessen?<br>[Bitte geben Sie an, wie Sie die Leistung in Bezug auf das in Checkliste 2 – Frage 2.1. angegebene Hauptziel messen.] ▶ |

| *Frage aus der Triage-Checkliste (Checkliste 1)* | *Antwort auf die Frage* | *Konsequenz für den Transparenzbericht (ausgehend von den Fragen in Checkliste 2)* |
|---|---|---|
| | | Beispiel: Wir haben die Leistungsfähigkeit des Algorithmus getestet, indem wir seine Vorhersagen auf Grundlage der historischen Daten überprüft haben, die von der Kantonspolizei zu Festnahmen nach Bewährung im Kanton erhoben wurden. Wir haben die Gesamtgenauigkeit gemessen, die Falsch-Negativ-Rate und die Falsch-Positiv-Rate. Der technische Anhang ist [hier] verfügbar.<br>2.8.1. Welche Methoden wurden verwendet, um die von den Systemvorhersagen/-empfehlungen/-entscheidungen unmittelbar betroffenen Stakeholder zu identifizieren? Und was sind die voraussichtlichen Auswirkungen auf diese Personen?<br>Antwort: siehe oben (Punkt 2.8.1. bei Frage 1.1.)<br>2.9. Welche Verfahren sind vorhanden, um Systemfehlern und Fehlfunktionen zu begegnen?<br>Beispiel: «Fehlerhafte Vorhersagen bezüglich Personen, die aus dem Gefängnis entlassen wurden und als risikoarm eingestuft wurden, aber erneut straffällig geworden sind, werden durch ein spezifisches Verfahren erfasst, das den Erfolg des zu einer Bewährungsstrafe verurteilten Straftäters überwacht. Alle Daten werden sicher in X [Angabe der Datenbanken oder des Registers] gespeichert und sind für Y [Rollen der öffentlichen Verwaltung] zugänglich. Die Maßnahme zur Behebung der fehlerhaften Vorhersagen lautet: Das Strafrecht sieht bereits rechtliche Konsequenzen für einen Rückfall während des Bewährungszeitraums vor. Der Anteil der Angeklagten, denen eine Bewährung gewährt oder abgelehnt wurde, wird erfasst und X, Y [Rollen in der Organisation] zugänglich gemacht. Die Möglichkeit falscher Prognosen in dem Sinne, dass ihre Risikobewertung auf Datenkorrekturfehlern basiert, wurde berücksichtigt und der folgende Plan wurde erstellt, um diese Art von Fehler zu minimieren [Verfahren zur Fehlerverminderung einfügen, falls vorhanden].» |
| **1.4. Wird das System verwendet, um eine Entscheidung über eine gesetzliche Pflicht oder ein Recht einer Person zu treffen?** | Ja, das System wird verwendet, um zu entscheiden, ob jemand das Gefängnis verlassen darf. | |

| *Frage aus der Triage-Checkliste (Checkliste 1)* | *Antwort auf die Frage* | *Konsequenz für den Transparenzbericht (ausgehend von den Fragen in Checkliste 2)* |
|---|---|---|
| **1.6. Können Einzelpersonen die Entscheidung durch Swiss COMPAS vermeiden oder verlangen, dass die Entscheidung mithilfe eines anderen Verfahrens getroffen wird, bei dem nicht dasselbe technische System verwendet wird?** | Zwei möglich Szenarien:<br>Ja, Angeklagte können es vermeiden, von *Swiss COMPAS* beurteilt zu werden, indem sie dies über ihren Anwalt verlangen, und dies wird immer akzeptiert. Infolgedessen wird die Rückfallgefahr von einem Richter ohne die Hilfe des Tools bewertet.<br>Nein, jeder Angeklagte muss die Bewertung akzeptieren. | Fall A) weiter zu Frage 1.10<br>Fall B) weiter zu Frage 1.7. |
| **1.7. Kann die Person, über die mithilfe des Tools eine Entscheidung getroffen wurde, beweisen, dass eine falsche Entscheidung getroffen wurde, ohne vor Gericht zu gehen?** | Nein, es kann nicht gezeigt werden, dass die Entscheidung, die Aussetzung zur Bewährung zu verweigern, auf einer falschen Prognose beruht. Der Nachweis, dass jemand, der auf Bewährung freigelassen wurde, tatsächlich erneut straffällig geworden ist, zeigt nicht, dass die Entscheidung *falsch* war, da die Entscheidung ausdrücklich auf einer unsicheren Risikobewertung beruhte, die (manchmal) damit vereinbar ist, dass ein Angeklagter mit geringem Rückfallrisiko erneut gegen das Gesetz verstößt. | 2.5. Wer ist verantwortlich dafür, Rückmeldungen der Endbenutzerinnen/Endbenutzer zu bearbeiten, d. h. der Personen, die das System benutzen oder von ihm unterstützt werden?<br>Beispiel: Richterinnen und Richter können ein Informationsgespräch mit Expertinnen und Experten der Firma *Swiss COMPAS* anfordern, die in Laienform erklären können, was das Tool berücksichtigt und warum oder welche Art von *Bias* im System vorliegt.<br>2.6. Wer ist dafür verantwortlich, auf Zweifel oder Anfechtungen durch Einzelne zu antworten, die von der Nutzung des Systems betroffen sind?<br>Beispiel: Niemand.<br>2.14. Wird der Einsatz des Systems nach der Testphase kontinuierlich überwacht?<br>a) Zu jedem Zeitpunkt?<br>b) In einem bestimmten Zeitraum?<br>c) Mit welchen Maßnahmen?<br>Beispiel: Nein. Es ist kein Überwachungssystem vorhanden, mit dem die Leistung des Systems über die Dauer seines Einsatzes verfolgt wird.<br>2.15. Können Personen, die von einer Entscheidung betroffen sind, den Output des automatisierten Systems erfahren und die vom System beeinflussten Vorhersagen/Empfehlungen/Entscheidungen anfechten? ▶ |

| *Frage aus der Triage-Checkliste (Checkliste 1)* | *Antwort auf die Frage* | *Konsequenz für den Transparenzbericht (ausgehend von den Fragen in Checkliste 2)* |
|---|---|---|
| | | Beispiel: Nein, es ist kein solches System entwickelt und implementiert worden.<br>2.17. Wie schneidet das System im Vergleich zu dem zuvor vorhandenen ab, falls es eines gibt, oder mit etablierten Benchmarks, falls vorhanden?<br>Beispiel: Das zuvor bestehende System war eine Bewährungsentscheidung menschlicher Richter unter Berücksichtigung von X, Y, Z. Bislang gibt es keine verlässlichen Studien zur Gesamtgenauigkeit von Bewährungsentscheidungen menschlicher Schweizer Richterinnen und Richter (Anteil der auf Bewährung freigelassenen Personen, die erneut straffällig werden). Es gibt einige Studien in den USA, die zeigen, dass US-Prognosetools x % genauer sind als menschliche Richter, aber es ist unklar, wie sich dies auf den Schweizer Kontext übertragen lässt, da die Leistung der Schweizer Richter nicht bekannt ist. Um die Genauigkeit von *Swiss COMPAS* zu messen, gehen wir davon aus, dass für jeden Angeklagten mit einem *Swiss-COMPAS*-Risikowert von mehr als 0,3 unabhängig vom Geschlecht eine Empfehlung zur Verweigerung der Bewährung angenommen wird (d. h., dass, wenn dieser Schwellenwert verwendet wird, von 1.000 Personen, die auf Bewährung freigelassen wurden, ungefähr 300 wieder straffällig geworden sind). Mit diesem Schwellenwert hat das *Swiss COMPAS* eine Genauigkeit von x %, was niedriger ist als die Genauigkeit des US-amerikanischen *COMPAS-Tools* mit y %. Die Falsch-Positiv-Rate (Anteil der Personen, bei denen in Testdaten ein erneuter Gesetzesverstoß vorhergesagt wurde, die aber nicht erneut straffällig geworden sind) ist bei *Swiss COMPAS* jedoch viel niedriger als beim US-*COMPAS*, während die Falsch-Negativ-Rate etwas höher liegt. Daher ist *Swiss COMPAS* besser als das US-amerikanische Gegenstück in der Lage, Personen zu identifizieren, die nicht erneut straffällig werden und eine Bewährung verdienen, aber es ist schlechter darin, Personen zu identifizieren, die erneut straffällig werden und deshalb im Gefängnis bleiben sollten. ▶ |

| *Frage aus der Triage-Checkliste (Checkliste 1)* | *Antwort auf die Frage* | *Konsequenz für den Transparenzbericht (ausgehend von den Fragen in Checkliste 2)* |
|---|---|---|
| | | 2.20. Wurden während der Überwachung Vorhersagen/ Empfehlungen/Entscheidungen des Systems jemals<br>a) vom Systemendbenutzer oder<br>b) von Personen, die Entscheidungen unterliegen, hinterfragt?<br>Beispiel: Das System ist in der Schweiz noch nicht in Betrieb. Es ist möglich, dass ähnliche Systeme in anderen Ländern, in denen sie eingesetzt werden, angefochten wurden, aber unsere Abteilung verfügt nicht über diese Informationen. |
| **1.10. Betrifft die Entscheidung einen der folgenden Bereiche des öffentlichen Lebens oder Ressourcen des öffentlichen Sektors:**<br>**• die Rechtspflege,**<br>**• den Zugang zu Bildungschancen,**<br>**• den Zugang zu demokratischen Prozessen**<br>**• (etc.)?** | Ja, es wird in der Justizverwaltung (Rechtspflege) verwendet. | 2.8.1. Welche Methoden wurden verwendet, um die von den Systemvorhersagen / -empfehlungen / -entscheidungen unmittelbar betroffenen Stakeholder zu identifizieren? Und was sind die voraussichtlichen Auswirkungen auf diese Personen?<br>Antwort: siehe oben (Punkt 2.8.1. bei Frage 1.1.)<br>2.20. Wurden während der Überwachung Vorhersagen/ Empfehlungen/Entscheidungen des Systems jemals<br>a) vom Systemendbenutzer oder<br>b) von Personen, die Entscheidungen unterliegen, hinterfragt?<br>Antwort: siehe oben (Punkt 2.20. bei Frage 1.7.) |

| *Frage aus der Triage-Checkliste (Checkliste 1)* | *Antwort auf die Frage* | *Konsequenz für den Transparenzbericht (ausgehend von den Fragen in Checkliste 2)* |
|---|---|---|
| **1.11. Findet durch die Anschaffung bzw. den Einsatz des KI-Systems in einem der folgenden Bereiche eine Änderung statt, bei öffentlicher Computer-Infrastruktur, öffentlichen Datenbeständen oder immateriellen Vermögenswerten (z. B. Kompetenzen) im öffentlichen Sektor?** | Ja. | 2.2.3. Welche Anforderungen werden an das System in Bezug auf die *Fairness* gestellt?<br>Antwort: siehe oben (Punkt 2.2.3. bei Frage 1.3.)<br>2.8.2. Welche Methoden wurden verwendet, um die von der digitalen Transformation in der öffentlichen Verwaltung betroffenen Personen zu identifizieren (z. B. Personal der öffentlichen Verwaltung)? Und was sind die voraussichtlichen Auswirkungen auf diese Personen?<br>Beispiel: Die Abteilungsleiter X, Y, Z des zuständigen Gerichts des Kantons Zürich haben sich getroffen und folgende Stakeholder und Konsequenzen identifiziert:<br>A) Richter: Die Verwendung des Algorithmus von *Swiss COMPAS* wird von den meisten Haftprüfungsrichtern aus zwei Gründen sehr begrüßt. Sie haben das Gefühl, dass sie angesichts der hohen Anzahl von Fällen, bei denen eine Haftprüfung notwendig ist, nicht genügend Zeit haben, um die Profile der Angeklagten bei der Entscheidung über eine Aussetzung zur Bewährung zu überprüfen, und Druck besteht, mehr Zeit für Folgeentscheidungen sowie die Aburteilung eines Verbrechens aufzuwenden; zum Teil aufgrund der kurzen Zeit, die sie dieser Aufgabe widmen sollen, beklagen sie sich über die Subjektivität und die schlechte Genauigkeit ihrer Einschätzungen und hoffen, dass die Genauigkeit ihrer Prognosen sowohl verbessert als auch objektiver oder zumindest uniform gestaltet werden kann. Basierend auf einer internen Umfrage ist es nur einer Minderheit solcher Richter wichtig, dass die Logik hinter der Risikobewertung für sie zugänglich gemacht wird, solange es starke Beweise dafür gibt, dass das Tool korrekt arbeitet<br>B) Kantonale Pflichtanwälte: Die kantonalen Pflichtanwälte (-verteidiger) begrüßen diesen Schritt nicht. Sie beschweren sich über die Undurchsichtigkeit des Tools, die es ihnen unmöglich macht, die Rechte der Menschen zu verteidigen, denen sie helfen. Sie planen den Gang zum obersten Gericht, wenn der Algorithmus hinter dem Score nicht zugänglich gemacht wird. |
| **1.12 Besteht das Risiko, dass das System eine politische Entscheidung (z. B. Wahl oder Volksabstimmung) beeinflusst?** | Nein. | |

| *Frage aus der Triage-Checkliste (Checkliste 1)* | *Antwort auf die Frage* | *Konsequenz für den Transparenzbericht (ausgehend von den Fragen in Checkliste 2)* |
|---|---|---|
| **1.13. Beeinflusst das technische System die Verteilung öffentlicher Mittel an wirtschaftliche Akteure in der Gesellschaft?** | Nein. | |
| **1.14. Beruht das technische System auf einem statistischen Modell des menschlichen Verhaltens oder der persönlichen Merkmale?** | Ja. | 2.2.3. Welche Anforderungen werden an das System in Bezug auf die *Fairness* gestellt?<br>[Hier werden die Anforderungen an das System in Sachen Fairness erläutert. Dieser Punkt ist zu komplex, um hier zusammengefasst zu werden – idealerweise sollte dies nach einer Konsultation mit Expertinnen und Experten und betroffenen Interessenvertreterinnen und -vertretern oder Strafrechtswissenschaftlerinnen und -wissenschaftlern definiert werden.]<br>Antwort: siehe oben (Punkt 2.2.3. bei Frage 1.3.)<br>2.8.1. Welche Methoden wurden verwendet, um die von den Systemvorhersagen/-empfehlungen/-entscheidungen unmittelbar betroffenen Stakeholder zu identifizieren? Und was sind die voraussichtlichen Auswirkungen auf diese Personen?<br>Antwort: siehe oben (Punkt 2.8.1. bei Frage 1.1.)<br>2.12. Welche Methoden wurden verwendet, um die Voreingenommenheit und die Fairness des Systems zu definieren und zu messen?<br>Beispiel: «Wir haben die Falscherkennungsrate, die Falschauslassungsrate, die Falsch-Positiv-Rate und die Falsch-Negativ-Rate insgesamt und nach Geschlecht aufgeschlüsselt gemessen, unter der Annahme, dass die Richter jedem Angeklagten mit einer Risikobewertung von mehr als 0,3 die Bewährung verweigern.<br>Die Quote falscher Auslassungen und falscher Entdeckungen ist jedoch für beide Geschlechter gleich. Risikobewertungen haben unabhängig vom Geschlecht, für das sie verwendet werden, den gleichen Prognosewert (sie bieten einen gleich starken Beleg dafür, dass ein Straftäter, dessen Haftstrafe zur Bewährung ausgesetzt wird, wieder straffällig wird). ▶ |

| *Frage aus der Triage-Checkliste (Checkliste 1)* | *Antwort auf die Frage* | *Konsequenz für den Transparenzbericht (ausgehend von den Fragen in Checkliste 2)* |
|---|---|---|
| | | Das System ist nicht adaptiv ausgelegt, da ein adaptives System für Personen mit denselben Merkmalen kein homogenes Ergebnis ergeben würde, d. h., Personen mit denselben Merkmalen erhalten möglicherweise unterschiedliche Empfehlungen, und das in einer Weise, die von den Richtern nur schwer zu kontrollieren wäre.»<br><br>2.13. Wie werden den Systemendbenutzern und den Personen, die vom Einsatz des Systems unmittelbar betroffen sind, individuelle Vorhersagen/Empfehlungen/Entscheidungen des Systems erklärt?<br><br>Beispiel: Die Richter wurden über die Merkmale informiert, die vom Vorhersagetool berücksichtigt werden. Die Formel und die spezifische Gewichtung dieser Merkmale sind jedoch ein Geschäftsgeheimnis und wurden ihnen aus diesem Grund nicht offenbart.<br><br>2.19. Bitte beschreiben Sie relevante Probleme mit *Bias*, die nicht gelöst werden konnten, oder mögliche Ursachen für Ungerechtigkeiten im System und erklären Sie, warum sie nicht gelöst werden können (beispielsweise, indem Sie Kompromisse mit anderen Systemzielen einschließlich widersprüchlicher Fairnessziele erläutern).<br><br>Beispiel: Es ist unmöglich, alle oben genannten Maße für alle Gruppen auszugleichen, da die durchschnittliche Rückfallwahrscheinlichkeit bei Männern und Frauen unterschiedlich ist. Für diesen Schwellenwert unterscheiden sich die Falsch-Positiv-Rate und die Falsch-Negativ-Rate für die Geschlechter.<br><br>Weibliche Gefangene haben eine niedrigere Falsch-Negativ-Rate, aber eine höhere Falsch-Positiv-Rate, d. h., es ist im Vergleich zu Männern weniger wahrscheinlich, dass sie zu Unrecht inhaftiert werden, und es ist wahrscheinlicher, dass sie zu Unrecht freigelassen.<br><br>Männliche Gefangene haben eine höhere Falsch-Positiv-Rate und eine niedrigere Falsch-Negativ-Rate, d. h., es ist wahrscheinlicher, dass sie fälschlicherweise im Gefängnis festgehalten werden, und weniger wahrscheinlich, dass sie fälschlicherweise inhaftiert werden.<br><br>Angesichts der in 2.2.3 angegebenen Fairnessziele und der Bedeutung des Ausgleichs der Rate falscher Entdeckungen und falscher Auslassungen (sowie der Übermittlung kalibrierter Bewertungen an die Richter), die sich aus dieser Analyse der Ziele des Systems ergibt, gehen wir davon aus, dass dies aus Sicht der Fairness die beste Lösung ist. ▶ |

| *Frage aus der Triage-Checkliste (Checkliste 1)* | *Antwort auf die Frage* | *Konsequenz für den Transparenzbericht (ausgehend von den Fragen in Checkliste 2)* |
|---|---|---|
| | | 2.20. Wurden während der Überwachung Vorhersagen/ Empfehlungen/Entscheidungen des Systems jemals<br>a) vom Systemendbenutzer oder<br>b) von Personen, die Entscheidungen unterliegen, hinterfragt?<br>Antwort: siehe oben (Punkt 2.20. bei Frage 1.7.) |
| **1.15. Ist das System so konzipiert, dass es adaptiv ist, sodass nicht alle neuen Fälle wie andere behandelt werden, denen es in der Vergangenheit begegnet ist, weil es seine Parameter ändert, z. B., um effizienter zu werden?** | Nein. | |
| **1.16. Ist es das Ziel des technischen Systems, ein vollständig deterministisches Regelsystem zu automatisieren, das nur ein Minimum an Kreativität und menschlichem Urteilsvermögen durch die derzeitigen menschlichen Anwenderinnen/ Anwender erfordert und keine Risiko- oder Wahrscheinlichkeitsabschätzungen beinhaltet?** | Nein. | |

| *Frage aus der Triage-Checkliste (Checkliste 1)* | *Antwort auf die Frage* | *Konsequenz für den Transparenzbericht (ausgehend von den Fragen in Checkliste 2)* |
|---|---|---|
| **1.17. Beruht das technische System auf Parametern, Merkmalen, Faktoren oder Entscheidungskriterien, die nicht dem entsprechen, was von den meisten Fachleuten auf diesem Gebiet normalerweise berücksichtigt wird?** | Unter der Annahme, dass «ein Experte auf dem Gebiet» = «Richter» ist, ist die Antwort Ja, da das System auf Merkmalen und Faktoren beruht, die von menschlichen Richtern normalerweise nicht (oder zumindest nicht systematisch) berücksichtigt werden. | 2.2.4. Welche Anforderungen werden an das System in Bezug auf die *Erklärbarkeit* gestellt?<br>Antwort: siehe oben (Punkt 2.2.4. bei Frage 1.3.)<br>2.8.1. Welche Methoden wurden verwendet, um die von den Systemvorhersagen/-empfehlungen/-entscheidungen unmittelbar betroffenen Stakeholder zu identifizieren? Und was sind die voraussichtlichen Auswirkungen auf diese Personen?<br>Antwort: siehe oben (Punkt 2.8.1. bei Frage 1.1.)<br>2.13. Wie werden den Systemendbenutzern und den Personen, die vom Einsatz des Systems unmittelbar betroffen sind, individuelle Vorhersagen/Empfehlungen/Entscheidungen des Systems erklärt?<br>Antwort: siehe oben (Punkt 2.12. bei Frage 1.15.)<br>2.14. Wird die Systembereitstellung nach der Testphase kontinuierlich überwacht?<br>a) Zu jedem Zeitpunkt?<br>b) In einem bestimmten Zeitrahmen?<br>c) Mit welchen Maßnahmen?<br>Antwort: siehe oben (Punkt 2.14. bei Frage 1.7.)<br>2.20. Wurden während der Überwachung Vorhersagen/Empfehlungen/Entscheidungen des Systems jemals<br>a) vom Systemendbenutzer oder<br>b) von Personen, die Entscheidungen unterliegen, hinterfragt?<br>Antwort: siehe oben (Punkt 2.20. bei Frage 1.7.) |
| **1.19. Greift das technische System auf die Infrastruktur eines Drittanbieters zurück, über die die öffentliche Einrichtung keine uneingeschränkte Kontrolle hat und/oder bei der sie keinen Zugriff auf z. B. Datensätze oder die Rechenleistung hat?** | Ja, der Algorithmus zur Erstellung des Risiko-Scores ist ein Geschäftsgeheimnis der Firma *Swiss COMPAS*. | 2.1. Für welches Problem soll das System eine Lösung liefern<br>Beispiel: Derzeit gibt es Zweifel an der Qualität von Bewährungsentscheidungen durch Richterinnen und Richter, die auch damit zusammenhängen, dass es ein Missverhältnis gibt zwischen der geringen Zeit, die den Richtern zur Verfügung steht, um diese Entscheidungen zu treffen, und der Anzahl der Fälle, in denen sie entscheiden müssen. *Swiss COMPAS* soll Richterinnen und Richtern eine Risikoanzeige in Verbindung mit einer optimierten Empfehlung (Bewährung zu gewähren oder zu verweigern) geben, die sie allerdings auch ignorieren dürfen. Das Ziel des Systems ist, bei der Prognose von Rückfällen sowohl genauer als auch homogener zu sein als die derzeitige Einschätzung durch menschliche Richterinnen und Richter. ▶ |

| *Frage aus der Triage-Checkliste (Checkliste 1)* | *Antwort auf die Frage* | *Konsequenz für den Transparenzbericht (ausgehend von den Fragen in Checkliste 2)* |
|---|---|---|
| | | 2.3. Wer ist für die Konstruktion des Systems verantwortlich?<br>Beispiel: Die Firma «Southpointe» mit Hauptsitz in Lugano, CH, in der Person von Gianni Contabene, CEO.<br>2.4. Wer ist für die Implementierung des Systems verantwortlich?<br>[Geben Sie die Rollen an, z. B. im kantonalen Strafgericht, die für die Haftprüfungsverfahren zuständig sind.]<br>2.5. Wer ist verantwortlich dafür, Rückmeldungen der Endbenutzerinnen/Endbenutzer zu bearbeiten, d. h. der Personen, die das System benutzen oder von ihm unterstützt werden?<br>Antwort: siehe oben (Punkt 2.5 bei Frage 1.7.)<br>2.6. Wer ist dafür verantwortlich, auf Zweifel oder Anfechtungen durch Einzelne zu antworten, die von der Nutzung des Systems betroffen sind?<br>Antwort: siehe oben (Punkt 2.6 bei Frage 1.7)<br>2.14. Wird die Systembereitstellung nach der Testphase kontinuierlich überwacht?<br>a) Zu jedem Zeitpunkt?<br>b) In einem bestimmten Zeitrahmen?<br>c) Mit welchen Maßnahmen?<br>Antwort: siehe oben (Punkt 2.14. bei Frage 1.7.)<br>2.15. Können Personen, die von einer Entscheidung betroffen sind, den Output des automatisierten Systems erfahren und die vom System beeinflussten Vorhersagen/Empfehlungen/Entscheidungen anfechten?<br>Antwort: siehe oben (Punkt 2.15. bei Frage 1.7.)<br>2.16. Wie verhält sich das System in Bezug auf die ausgewählten relevanten Metriken?<br>[Bitte beachten Sie alle Ziele und Anforderungen, die in Checkliste 2 – Fragen 2.1. angegeben sind.]<br>Das erste Ziel ist, dass die Vorhersage so genau wie möglich ist: Es soll die Anzahl der Personen minimiert werden, denen die Bewährung verweigert wird, obwohl sie nicht wieder straffällig würden, und derjenigen, die auf Bewährung entlassen werden, aber wieder eine Straftat begehen würden. |

| *Frage aus der Triage-Checkliste (Checkliste 1)* | *Antwort auf die Frage* | *Konsequenz für den Transparenzbericht (ausgehend von den Fragen in Checkliste 2)* |
|---|---|---|
| | | Das System hat eine Gesamtgenauigkeit von x %. Der Vergleich mit relevanten nationalen und internationalen Benchmarks wird weiter oben diskutiert, siehe Abschnitt 2.17.<br>2.17. Wie schneidet das System im Vergleich zu dem zuvor vorhandenen ab, falls es eines gibt, oder mit etablierten Benchmarks, falls vorhanden?<br>Antwort: siehe oben (Punkt 2.17. bei Frage 1.7.)<br>2.19. Bitte beschreiben Sie relevante Probleme mit *Bias*, die nicht gelöst werden konnten, oder mögliche Ursachen für Ungerechtigkeiten im System und erklären Sie, warum sie nicht gelöst werden können (beispielsweise, indem Sie Kompromisse mit anderen Systemzielen einschließlich widersprüchlicher Fairnessziele erläutern).<br>Antwort: siehe oben (Punkt 2.19. bei Frage 1.13.)<br>2.8.2. Welche Methoden wurden verwendet, um die von der digitalen Transformation in der öffentlichen Verwaltung betroffenen Personen zu identifizieren (z. B. Personal der öffentlichen Verwaltung)? Und was sind die voraussichtlichen Auswirkungen auf diese Personen?<br>Antwort: siehe oben (Punkt 2.8.2. bei Frage 1.11.)<br>2.20. Wurden während der Überwachung Vorhersagen/Empfehlungen/Entscheidungen des Systems jemals hinterfragt? |
| **Überprüfung: Resultiert aus der Beantwortung der Fragen in Checkliste 1, dass Sie einen Transparenzbericht schreiben sollen?** | Ja. | Siehe Antworten bei 1.19 |
| **Letzte Nachfrage: Gibt es zusätzliche ethische Fragen?** | Nein. | Beispiel: Zusätzliche ethische Fragen sind uns nicht bekannt. |

## 3. Transparenzbericht

Es folgt der Transparenzbericht, der sich aus der Beantwortung der Checkliste 1 für diesen hypothetischen Fall (*Swiss COMPAS*) ergibt. Er weist Antworten auf alle Fragen auf, doch das ist üblicherweise nicht zu erwarten. Der Transparenzbericht soll Antworten enthalten, die durch die Fragen in Checkliste 1 notwendig werden. Dass hier alle Fragen der Checkliste beantwortet werden, dient lediglich der Veranschaulichung.

### a) Was soll das System leisten und welchen Anforderungen an den Schutz von Grundwerten soll es gerecht werden?

**2.1.** Für welches Problem soll das System eine Lösung liefern?

Beispiel: Derzeit gibt es Zweifel an der Qualität von Bewährungsentscheidungen durch Richterinnen und Richter, die auch damit zusammenhängen, dass es ein Missverhältnis zwischen der geringen Zeit, die den Richtern zur Verfügung steht, um diese Entscheidungen zu treffen, und der Anzahl der Fälle, in denen sie entscheiden müssen, gibt. *Swiss COMPAS* soll Richterinnen und Richtern eine Risikoanzeige in Verbindung mit einer optimierten Empfehlung (Bewährung zu gewähren oder zu verweigern) geben, die sie allerdings auch ignorieren dürfen. Das Ziel des Systems ist, bei der Prognose von Rückfällen sowohl genauer als auch homogener zu sein als die derzeitige Einschätzung durch menschliche Richterinnen und Richter.

**2.2.** Weitere Anforderungen des Systems?

**2.2.1.** Welche Anforderungen werden an das System in Bezug auf die *Privatsphäre* gestellt?

[Legen Sie hier die Anforderungen an das System dar. Welche Maßnahmen bezüglich des Datenschutzes sollten ergriffen werden? Nehmen Sie beispielsweise Kontakt mit der/dem Datenschutzbeauftragten Ihrer Organisation auf, um diesen Teil des Berichts zu verfassen.]

**2.2.2.** Welche Anforderungen werden an das System in Bezug auf die *Cybersicherheit* gestellt?

[Legen Sie hier die Anforderungen an das System dar. Fordern Sie von Cybersicherheitsexperten eine technische Expertise an.]

**2.2.3.** Welche Anforderungen werden an das System in Bezug auf die *Fairness* gestellt?

[Dieser Aspekt ist zu komplex, um hier darauf einzugehen – die Beurteilung erfordert eine gemeinsame Analyse von zumindest Expertinnen und Experten für Statistik und Kriminologie, die in der Lage sind, eine begründete Einschätzung dazu zu geben, was sie für ein faires und unvoreingenommenes Urteil halten; im Idealfall sollten Strafverteidiger und/oder Experten für soziale Gerechtigkeit beauftragt werden oder die Ergebnisse überprüfen.]

Beispiel: Wir definieren Fairness wie folgt: Die Vorhersage, ob jemand erneut straffällig wird, ist im Durchschnitt für Männer und Frauen gleichermaßen genau.

Die Rechtfertigung dafür, Fairness so zu bestimmen, ist: …

**2.2.4.** Welche Anforderungen werden an das System in Bezug auf die *Erklärbarkeit* gestellt?

Beispielhafte Erwägungen: Da dies eine weitreichende Entscheidung ist, die letztendlich von Menschen (Richterinnen und Richtern) getroffen wird, erscheint es wichtig, dass sie ein mentales Modell der Faktoren und ihrer Gewichtung bilden können, die bei der Erzeugung des Scores berücksichtigt wurden. Sofern das Schweizer *COMPAS*-Tool einen geheimen Algorithmus verwendet, kann diese Anforderung möglicherweise nicht erfüllt werden. Wenn die Formel jedoch öffentlich bekannt wäre, würden um Bewährung ersuchende Personen unaufrichtig antworten, um ihre Risikobewertung zu verbessern. (Der US-*COMPAS*-Algorithmus ist geheim.)

Ein realer Bericht würde im Idealfall eine gründlichere Analyse und Diskussion dieses Widerspruchs sowie potenziell realisierbare Lösungen beinhalten.

## b) Wer ist rechenschaftspflichtig?

**2.3.** Wer ist für die Konstruktion des Systems verantwortlich?

Beispiel: Die Firma «Southpointe» mit Hauptsitz in Lugano, CH, in der Person von Gianni Contabene, CEO.

**2.4.** Wer ist für die Implementierung des Systems verantwortlich?

[Geben Sie die Rollen an, z. B. im kantonalen Strafgericht, die für die Haftprüfungsverfahren zuständig sind.]

**2.5.** Wer ist verantwortlich dafür, Rückmeldungen der Endbenutzerinnen/Endbenutzer zu bearbeiten, d. h. der Personen, die das System benutzen oder von ihm unterstützt werden?

Beispiel: Richterinnen und Richter können ein Informationsgespräch mit Expertinnen und Experten der Firma *Swiss COMPAS* anfordern, die in Laienform erklären können, was das Tool berücksichtigt und warum oder welche Art von *Bias* im System vorliegt.

**2.6.** Wer ist dafür verantwortlich, auf Zweifel oder Anfechtungen durch Einzelne zu antworten, die von der Nutzung des Systems betroffen sind?

Beispiel: Niemand.

## c) Transparenzinformationen über die Umsetzung und Steuerung des Systems

**2.7.** Mit welchen Methoden wurde die Leistung des Systems getestet und gemessen? [Bitte geben Sie an, wie Sie die Leistung in Bezug auf das in Checkliste 2 – Frage 2.1. angegebene Hauptziel messen.]

**2.8.** Welche Methoden wurden verwendet)

**2.8.1.** Welche Methoden wurden verwendet, um die von den Systemvorhersagen/-empfehlungen/-entscheidungen unmittelbar betroffenen Stakeholder zu identifizieren? Und was sind die voraussichtlichen Auswirkungen auf diese Personen?

Beispiel: «Wir haben ein Brainstorming-Meeting mit Staatsanwälten und Richtern des Kantons und Anwälten der Strafjustiz durchgeführt. Bei diesem Treffen wurden die Stakeholder identifiziert, die direkt von den Vorhersagen betroffen sind, nämlich die Angeklagten, ihre Verteidiger, ihre Familien, potenzielle künftige Opfer, wenn die Angeklagten erneut straffällig werden, und Gemeinschaften, in denen Menschen, die möglicherweise erneut straffällig werden, leben.

Unsere Analyse der Stakeholderinteressen sieht wie folgt aus:

A) Angeklagte. Das System zu verwenden liegt im Interesse der Angeklagten, bei denen es unwahrscheinlich ist, dass sie erneut straffällig werden (oder die statistisch nicht von denen zu unterscheiden sind, bei denen es unwahrscheinlich ist, dass sie erneut straffällig werden). Es ist insbesondere im Interesse derjenigen, die ihr Recht auef ein günstiges Bewährungsurteil am schlechtesten ausüben können, da sie sich nicht die besten Anwälte leisten können. Es ist nicht im Interesse von Personen, die sich mithilfe guter Anwälte bessere Chancen verschaffen können, auf Bewährung entlassen zu werden.

B) Anwälte der Angeklagten. Das System ist nicht in ihrem Interesse, da es ein Bestandteil richterlicher Entscheidung sein wird, den die Anwälte nicht anfechten können.

C) Familien. Die Familien der Angeklagten werden dann von der höheren Wahrscheinlichkeit profitieren, dass der Angeklagte auf Bewährung freigelassen wird, wenn der Einsatz von *Swiss COMPAS* im Vergleich zum Status quo zu einem höheren Anteil an gewährten Bewährungsstrafen führt (es sei denn, Angeklagte sind wegen eines Verbrechens gegen ihre Familien angeklagt). Dies hängt eng mit der Verhältnismäßigkeit der Entscheidungen zur Gewährung der Bewährung zusammen.

D) Potenzielle Opfer. Sie werden dann von *Swiss COMPAS* profitieren, wenn dadurch ein geringerer

Anteil der wieder straffällig werdenden Angeklagten freigelassen wird. Dieses Interesse wird nicht unbedingt gefördert, wenn *Swiss COMPAS* zu einem geringeren Anteil an gewährten Bewährungsstrafen führt. Wenn das Tool einerseits weniger genau ist als menschliche Richter, kann ein geringerer Anteil der gewährten Bewährungsstrafen dazu führen, dass die Quote der Straftäter, deren Haftstrafe zur Bewährung ausgesetzt wurde, die aber erneut straffällig werden, steigt. Wenn das Tool andererseits jedoch genauer ist als menschliche Richter, kann ein höherer Anteil der gewährten Bewährungsstrafen damit einhergehen, dass weniger Verbrechen von Straftätern begangen werden, deren Haftstrafe zur Bewährung ausgesetzt wurde.

E) Das Interesse der Gemeinschaften, die davon profitieren könnten, kann als Kombination folgender Interessen angesehen werden:

— der Familienmitglieder der Angeklagten, wie in Buchstabe C oben angegeben,

— der Personen, deren Interessen mit denen der Familienmitglieder in Einklang stehen,

— der Interessen der potenziellen Opfer von Straftätern, deren Haftstrafe zur Bewährung ausgesetzt wurde, wie oben in D angegeben,

— der Personen, deren Interessen mit denen der potenziellen Opfer von Straftätern in Einklang stehen, deren Haftstrafe zur Bewährung ausgesetzt wurde (z. B. der Kinder des Opfers).

Ein Tool, das in der Lage ist, mehr Menschen auf Bewährung freizulassen, während gleichzeitig die Häufigkeit der erneuten Straffälligkeit von Straftätern, deren Haftstrafe zur Bewährung ausgesetzt wurde, abnimmt, sollte von den Gemeinschaften der Angeklagten begrüßt werden.»

**2.8.2.** Welche Methoden wurden verwendet, um die von der digitalen Transformation in der öffentlichen Verwaltung betroffenen Personen zu identifizieren (z. B. Personal der öffentlichen Verwaltung)? Und was sind die voraussichtlichen Auswirkungen auf diese Personen?

Beispiel: Die Abteilungsleiter X, Y, Z des zuständigen Gerichts des Kantons Zürich haben sich getroffen und folgende Stakeholder und Konsequenzen identifiziert:

A) Richter: Die Verwendung des Algorithmus von *Swiss COMPAS* wird von den meisten Haftprüfungsrichtern aus zwei Gründen sehr begrüßt. Sie haben das Gefühl, dass sie angesichts der hohen Anzahl von Fällen, bei denen eine Haftprüfung notwendig ist, nicht genügend Zeit haben, um die Profile der Angeklagten bei der Entscheidung über eine Aussetzung zur Bewährung zu überprüfen, und Druck besteht, mehr Zeit für Folgeentscheidungen sowie die Aburteilung eines Verbrechens aufzuwenden; zum Teil aufgrund der kurzen Zeit, die sie dieser Aufgabe widmen sollen, beklagen sie sich über die Subjektivität und die schlechte Genauigkeit ihrer Einschätzungen und hoffen, dass die Genauigkeit ihrer Prognosen sowohl verbessert als auch objektiver oder zumindest uniform gestaltet werden kann. Basierend auf einer internen Umfrage ist es nur einer Minderheit solcher Richter wichtig, dass die Logik hinter der Risikobewertung für sie zugänglich gemacht wird, solange es starke Beweise dafür gibt, dass das Tool korrekt arbeitet.

B) Kantonale Pflichtanwälte: Die kantonalen Pflichtanwälte (-verteidiger) begrüßen diesen Schritt nicht. Sie beschweren sich über die Undurchsichtigkeit des Tools, die es ihnen unmöglich macht, die Rechte der Menschen zu verteidigen, denen sie helfen. Sie planen den Gang zum obersten Gericht, wenn der Algorithmus hinter dem Score nicht zugänglich gemacht wird.

**2.9.** Welche Verfahren sind vorhanden, um Systemfehler und Fehlfunktionen zu behandeln?

[Hier fügen Sie einen Abschnitt zur Cybersicherheit ein. Sie erläutern beispielsweise, wie Sie mit Fehlern von Mitarbeitern umgehen, die die Integrität, Verfügbarkeit oder Vertraulichkeit der gesammelten

Informationen gefährden. Dieser Abschnitt wird am besten von Cybersicherheitsexperten entworfen.]

**2.10.** Welche Methoden wurden zur Definition und zum Schutz der Privatsphäre verwendet?

[Bitte gehen Sie in diesem Teil des Berichts spezifisch auf die in Checkliste 1 – Frage 1.1. genannten Aspekte ein.]

**2.11.** Welche Maßnahmen zum Schutz der Cybersicherheit wurden getroffen?

[In diesem Abschnitt erläutern Sie die im System integrierten Cybersicherheitsmaßnahmen. Dieser Abschnitt wird am besten von Cybersicherheitsexperten entworfen.]

**2.12.** Welche Methoden wurden verwendet, um die Voreingenommenheit und die Fairness des Systems zu definieren und zu messen?

Beispiel: Wir haben die Falscherkennungsrate, die Falschauslassungsrate, die Falsch-Positiv-Rate und die Falsch-Negativ-Rate insgesamt und nach Geschlecht aufgeschlüsselt gemessen, unter der Annahme, dass die Richter jedem Angeklagten mit einer Risikobewertung von mehr als 0,3 die Bewährung verweigern.

Die Quote falscher Auslassungen und falscher Entdeckungen ist jedoch für beide Geschlechter gleich. Risikobewertungen haben unabhängig vom Geschlecht, für das sie verwendet werden, den gleichen Prognosewert (sie bieten einen gleich starken Beleg dafür, dass ein Straftäter, dessen Haftstrafe zur Bewährung ausgesetzt wird, wieder straffällig wird).

Das System ist nicht adaptiv ausgelegt, da ein adaptives System für Personen mit denselben Merkmalen kein homogenes Ergebnis ergeben würde, d. h., Personen mit denselben Merkmalen erhalten möglicherweise unterschiedliche Empfehlungen, und das in einer Weise, die von den Richtern nur schwer zu kontrollieren wäre.

**2.13.** Wie werden den Systemendbenutzern und den Personen, die vom Einsatz des Systems unmittelbar betroffen sind, individuelle Vorhersagen/Empfehlungen/Entscheidungen des Systems erklärt?

Beispiel: Die Richter wurden über die Merkmale informiert, die vom Vorhersagetool berücksichtigt werden. Die Formel und die spezifische Gewichtung dieser Merkmale sind jedoch ein Geschäftsgeheimnis und wurden ihnen aus diesem Grund nicht offenbart.

**2.14.** Wird die Systembereitstellung nach der Testphase kontinuierlich überwacht?

a) Zu jedem Zeitpunkt?

b) In einem bestimmten Zeitrahmen?

c) Mit welchen Maßnahmen?

Beispiel: Nein. Es ist kein Überwachungssystem vorhanden, mit dem die Leistung des Systems über die Dauer seines Einsatzes verfolgt wird.

**2.15.** Können Personen, die von einer Entscheidung betroffen sind, den Output des automatisierten Systems erfahren und die vom System beeinflussten Vorhersagen/Empfehlungen/Entscheidungen anfechten?

Beispiel: Nein, es ist kein solches System entwickelt und implementiert worden.

### d) Transparenzinformationen über die Leistungen des Systems

Auf der Grundlage der bisherigen Testläufe:

**2.16.** Wie verhält sich das System in Bezug auf die ausgewählten relevanten Metriken?

[Bitte beachten Sie alle Ziele und Anforderungen, die in Checkliste 2 – Fragen 2.1. angegeben sind.]

Beispiel: Das erste Ziel ist, dass die Vorhersage so genau wie möglich ist: Es soll die Anzahl der Personen minimiert werden, denen die Bewährung verweigert wird, obwohl sie nicht wieder straffällig würden, und

derjenigen, die auf Bewährung entlassen werden, aber wieder eine Straftat begehen würden.

Das System hat eine Gesamtgenauigkeit von x %. Der Vergleich mit relevanten nationalen und internationalen Benchmarks wird weiter oben diskutiert, siehe Abschnitt 2.17**.**

**2.17.** Wie schneidet das System im Vergleich zu dem zuvor vorhandenen ab, falls es eines gibt, oder mit etablierten Benchmarks, falls vorhanden?

Beispiel: Das zuvor bestehende System war eine Bewährungsentscheidung menschlicher Richter unter Berücksichtigung von X, Y, Z. Bislang gibt es keine verlässlichen Studien zur Gesamtgenauigkeit von Bewährungsentscheidungen menschlicher Schweizer Richterinnen und Richter (Anteil der auf Bewährung freigelassenen Personen, die erneut straffällig werden). Es gibt einige Studien in den USA, die zeigen, dass US-Prognosetools x % genauer sind als menschliche Richter, aber es ist unklar, wie sich dies auf den Schweizer Kontext übertragen lässt, da die Leistung der Schweizer Richter nicht bekannt ist. Um die Genauigkeit von *Swiss COMPAS* zu messen, gehen wir davon aus, dass für jeden Angeklagten mit einem *Swiss-COMPAS*-Risikowert von mehr als 0,3 unabhängig vom Geschlecht eine Empfehlung zur Verweigerung der Bewährung angenommen wird (d. h., dass, wenn dieser Schwellenwert verwendet wird, von 1.000 Personen, die auf Bewährung freigelassen wurden, ungefähr 300 wieder straffällig geworden sind). Mit diesem Schwellenwert hat das *Swiss COMPAS* eine Genauigkeit von x %, was niedriger ist als die Genauigkeit des US-amerikanischen *COMPAS*-Tools mit y %. Die Falsch-Positiv-Rate (Anteil der Personen, bei denen in Testdaten ein erneuter Gesetzesverstoß vorhergesagt wurde, die aber nicht erneut straffällig geworden sind) ist bei *Swiss COMPAS* jedoch viel niedriger als beim US-*COMPAS*, während die Falsch-Negativ-Rate etwas höher liegt. Daher ist *Swiss COMPAS* besser als das US-amerikanische Gegenstück in der Lage, Personen zu identifizieren, die nicht erneut straffällig werden und eine Bewährung verdienen, aber es ist schlechter darin, Personen zu identifizieren, die erneut straffällig werden und deshalb im Gefängnis bleiben sollten.

**2.18**. Welches sind die verbleibenden Sicherheits- und Datenschutzrisiken und warum sind sie angemessen?

[Hier erklären Sie, warum das Cybersicherheitsrisiko, das sich aus den Ziffern 2.9. und 2.11. ergibt, angesichts dessen, was auf dem Spiel steht, und der Wahrscheinlichkeit einer Attacke als angemessen beurteilt wird.]

**2.19.** Bitte beschreiben Sie relevante Probleme mit *Bias*, die nicht gelöst werden konnten, oder mögliche Ursachen für Ungerechtigkeiten im System und erklären Sie, warum sie nicht gelöst werden können (beispielsweise, indem Sie Kompromisse mit anderen Systemzielen einschließlich widersprüchlicher Fairnessziele erläutern).

Beispiel: Es ist unmöglich, alle oben genannten Maße für alle Gruppen auszugleichen, da die durchschnittliche Rückfallwahrscheinlichkeit bei Männern und Frauen unterschiedlich ist. Für diesen Schwellenwert unterscheiden sich die Falsch-Positiv-Rate und die Falsch-Negativ-Rate für die Geschlechter.

Weibliche Gefangene haben eine niedrigere Falsch-Negativ-Rate, aber eine höhere Falsch-Positiv-Rate, d. h. es ist im Vergleich zu Männern weniger wahrscheinlich, dass sie zu Unrecht inhaftiert werden, und es ist wahrscheinlicher, dass sie zu Unrecht freigelassen werden.

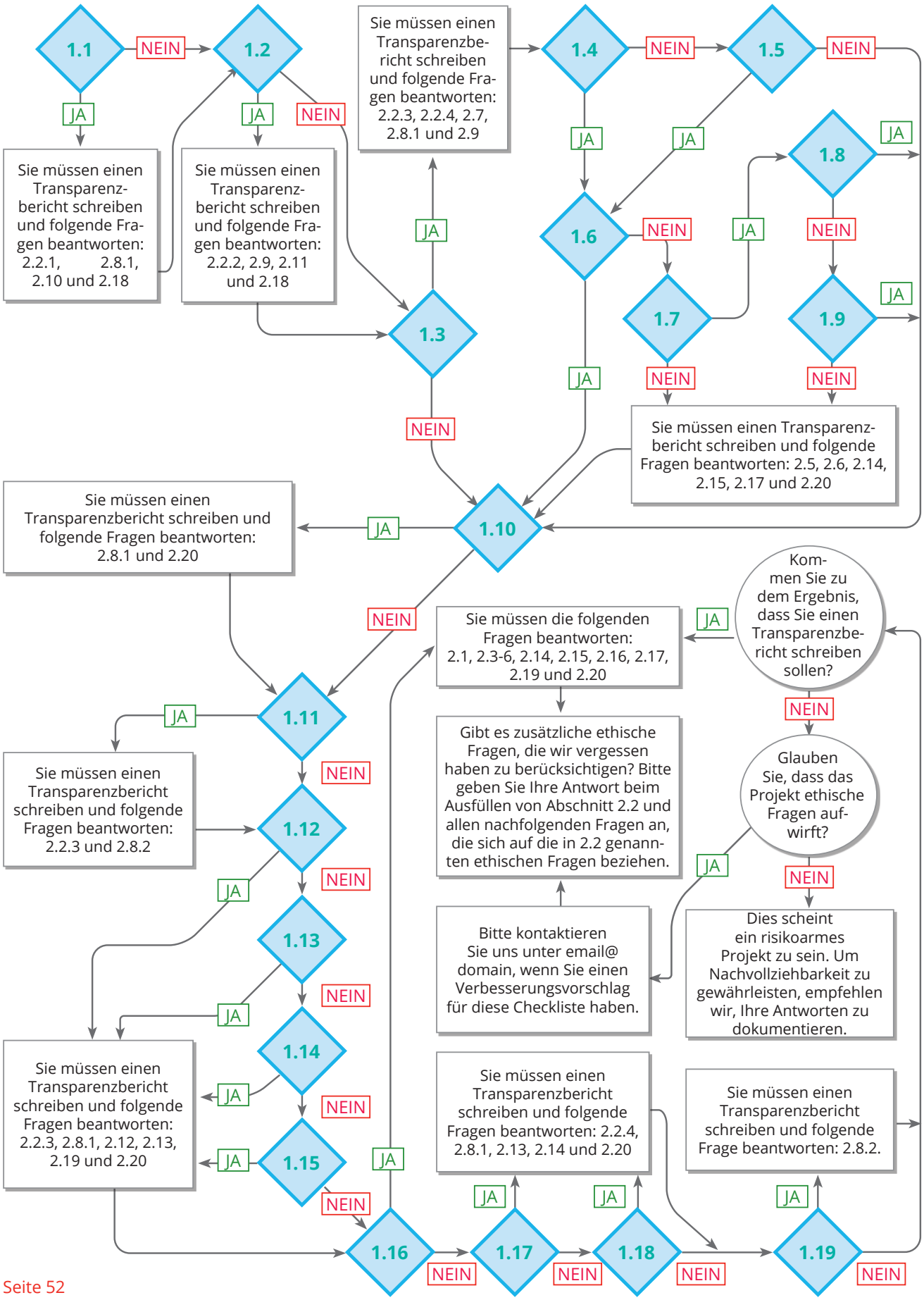
Männliche Gefangene haben eine höhere Falsch-Positiv-Rate und eine niedrigere Falsch-Negativ-Rate, d. h., es ist wahrscheinlicher, dass sie fälschlicherweise im Gefängnis festgehalten werden, und weniger wahrscheinlich, dass sie fälschlicherweise inhaftiert werden.

Angesichts der in 2.2.3 angegebenen Fairnessziele und der Bedeutung des Ausgleichs der Rate falscher Entdeckungen und falscher Auslassungen (sowie der Übermittlung kalibrierter Bewertungen an die Richter), die sich aus dieser Analyse der Ziele des Systems ergibt, gehen wir davon aus, dass dies aus Sicht der Fairness die beste Lösung ist.

**2.20.** Wurden während der Überwachung Vorhersagen/Empfehlungen/ Entscheidungen des Systems jemals hinterfragt?

Beispiel: Das System ist in der Schweiz noch nicht in Betrieb. Es ist möglich, dass ähnliche Systeme in anderen Ländern, in denen sie eingesetzt werden, angefochten wurden, aber unsere Abteilung verfügt nicht über diese Informationen.

## / Flussdiagramm



# / Literaturverzeichnis

Die nachstehenden Werke werden, wenn nichts anderes angegeben ist, mit Nachnamen des Autors, Jahreszahl sowie mit Seitenzahlen oder Randnummern zitiert.

**Albertini Michele** (2000), Der verfassungsmäßige Anspruch auf rechtliches Gehör im Verwaltungsverfahren des modernen Staates, Eine Untersuchung über Sinn und Gehalt der Garantie unter besonderer Berücksichtigung der bundesgerichtglichen Rechtsprechung, Diss. Bern

**Albus James** (1991), Outline for a theory of intelligence, in: Vol. 21, No. 3, IEEE Transactions on Systems, Man, and Cybernetics

**Aliotta Massimo** (2014), § 6 Medizinische Begutachtung, in: Steiger-Sackmann Sabine/ Mosimann Hans-Jakob (Hrsg.), Recht der Sozialen Sicherheit, Sozialversicherungen, Opferhilfe, Sozialhilfe – Beraten und Prozesszieren, Basel, S. 245 ff.

**Allhutter Doris/Mager Astrid/Cech Florian/ Fischer Fabian/Grill Gabriel** (2020), Der AMS-Algorithmus. Eine Soziotechnische Analyse des Arbeitsmartkchancen-Assistenz-Systems (AMAS), Institut für Technikfolgen-Abschätzung der Österreichischen Akademie der Wissenschaften, abrufbar unter: http://epub.oeaw.ac.at/0xc1aa5576_0x003bfdf3.pdf

**Althaus-Houriet Isabelle** (2017), Kommentierung der Art. 122–131 DBG, Noël Yves/Aubry Girardin Florence (Hrsg.), Commentaire Romand de la loi sur l'Impôt fédéral direct, 2. Aufl., Basel

**Auer Christoph/Müller Markus/Schindler Benjamin** (Hrsg.) (2018), VwVG, Kommentar zum Bundesgesetz über das Verwaltungsverfahren, 2. Aufl., Zürich (zit. Bearbeiter/in, 2018, N … zu Art. … VwVG)

**Baeriswyl Bruno/Pärli Kurt** (Hrsg.) (2015), Stämpfli Handkommentar zum Datenschutzgesetz, Bern, (zit. Bearbeiter/in, 2015, N … zu § … DSG)

**Baeriswyl Bruno/Rudin Beat** (Hrsg.) (2012), Praxiskommentar zum Informations- und Datenschutzgesetz des Kantons Zürich (IDG), Zürich/ Basel/Genf (zit. Bearbeiter/in, 2012, N … zu § … IDG)

**Beauchamp Tom L./Childress James F.** (2008), Principles of Biomedical Ethics, 6. Aufl., New York

**Beck Susanne /Grunwald Armin/Jacon Kai/ Matzner Tobias** (2019), Künstliche Intelligenz und Diskriminierung, Herausforderungen und Lösungsansätze, Whitepaper aus der Plattform Lernende Systeme, München, abrufbar unter: https://www.plattform-lernende-systeme.de/publikationen-details/kuenstliche-intelligenz-und-diskriminierung-herausforderungen-und-loesungsansaetze.html

**Biaggini Giovanni** (2017), Kommentar zur Bundesverfassung der schweizerischen Eidgenossenschaft, 2. Aufl., Zürich (zit. Biaggini, 2017, N … zu Art. … BV)

**Blumenstein Ernst/Locher Peter** (2016), System des schweizerischen Steuerrechts, 7. Aufl. von Professur Dr. Peter Locher, Zürich/Basel/Genf

**Bomhard David** (2019), Automatisierung und Entkollektivierung betrieblicher Arbeitsorganisation, Herausforderungen einer digitalen Arbeitsweilt, Diss. München 2018

**Braun Binder Nadja/Brändli Daniel** (2003), Vote électronique – Abstimmen und Wählen per Mausklick, in: LeGes 2003, S. 125 ff.

**Braun Binder Nadja** (2016a), Auf dem Weg zum vollautomatisierten Besteuerungsverfahren in Deutschland, in: Jusletter IT vom 25. Mai 2016

**Braun Binder Nadja** (2016b), Ausschließlich automationsgestützt erlassene Steuerbescheide und Bekanntgabe durch Bereitstellung zum Datenabruf, in: DStZ 2016, S. 526 ff.

**Braun Binder Nadja** (2016c), Vollständig automatisierter Erlass eines Verwaltungsaktes und Bekanntgabe über Behördenportale, in: DÖV 21/2016, S. 891 ff.

**Braun Binder Nadja** (2016d), Vollautomatisierte Verwaltungsverfahren im allgemeinen Veraltungsverfahrensrecht?, in: NvWZ 2016, S. 960 ff.

**Braun Binder Nadja** (2018), Algorithmic Regulation – Der Einsatz algorithmischer Verfahren im staatlichen Steuerungskontext, in: Hill Hermann/ Wieland Joachim (Hrsg.), Zukunft der Parlamente - Speyer Konvent in Berlin, S. 107 ff.

**Braun Binder Nadja** (2019a), Vollautomatisiert erlassene Verwaltungsakte und elektronische Aktenführung, in: Seckelmann Margrit (Hrsg.), Digitalisierte Verwaltung – Vernetztes E-Government, Berlin 2019, S. 311 ff.

**Braun Binder Nadja** (2019b), Künstliche Intelligenz und automatisierte Entscheidungen in der öffentlichen Verwaltung, in: SJZ 15/2019, S. 467 ff.

**Braun Binder Nadja** (2020a), Automatisierte Entscheidungen: Perspektive Datenschutzrecht und öffentliche Verwaltung, in: SZW 1/2020, S. 27 ff.

**Braun Binder Nadja** (2020b), Als Verfügung gelten Anordnungen der Maschinen im Einzelfall … – Dystopie oder künftiger Verwaltungsalltag?, in: ZSR 139/2020, S. 253 ff.

**Braun Binder Nadja** (2020c), Der Untersuchungsgrundsatz als Herausforderung vollautomatisierter Verfahren, in: zsis) 2/2020, A5, S. 26 ff.

**Braun Binder Nadja** (2020d), AI and Taxation: Risk Management in Fully Automated Taxation Procedures, in: Rademacher Timo/Wischmeyer Thomas (Hrsg.), Regulating Artificial Intelligence, Wiesbaden, S. 295 ff.

**Bryson Joanna** (2017), Three very different sources of bias in AI, and how to fix them, abrufbar unter: https://joanna-bryson.blogspot.com/2017/07/three-very-different-sources-of-bias-in.html

**Bürkle Martin** (2020), Kommentierungen zu Art. 1 und 2, in: Frésard-Fellay Ghislaine/Klett Barbara/ Leuzinger Susanne (Hrsg.), Basler Kommentar zum Allgemeinen Teil des Sozialversicherungsrechts, (zit. Bürkle, 2020, N … zu Art. … ATSG)

**Danaher John** (2016a), The Threat of Algocracy: Reality, Resistance and Accommodation, in: Philosophy & Technology, S. 1 ff., https://doi.org/10.1007/s13347-015-0211-1

**Danaher John** (2016b), Will Life Be Worth Living in a World Without Work? Technological Unemployment and the Meaning of Life, in: Science and Engineering Ethics, abrufbar unter: https://doi.org/10.1007/s11948-016-9770-5

**Dawson et al.** (2020), Artificial Intelligence: Australia's Ethics Framework, abrufbar unter: https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/supporting_documents/ArtificialIntelligenceethicsframeworkdiscussionpaper.pdf

**de Laat Paul B.** (2017), Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?, in: Philosophy & Technology, abrufbar unter: https://doi.org/10.1007/s13347-017-0293-z

**Degrandi Benno** (1977), Die automatisierte Verwaltungsverfügungen, Diss. Zürich

**Demaj Labinot/Sägesser Patrick** (2017), Chatbots für Organisationen: Anatomie, Potentiale und Anwendungsmöglichkeiten zur direkten Kommunikation mit Anspruchsgruppen, Diskussionspapier, abrufbar unter: https://byerley.ch/assets/pdf/bylerley_2017_Chatbots_Diskussionspapier.pdf

**Djeffal Christian** (2017), Das Internet der Dinge und die öffentliche Verwaltung – Auf dem Weg zum automatisierten Smart Government?, in: DVBl 2017, S. 808 ff.

**Djeffal, Christian** (2020), Künstliche Intelligenz, in: Klenk Tanja/Nullmeier Frank/Wewer Goettrik (Hrsg.): Handbuch Digitalisierung in Staat und Verwaltung, Wiesbaden

**Döring Nicola/Bortz Jürgen** (2016), Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften, 5. Aufl., Berlin

**Egli Catherine** (2020), Automatisierte Einzelentscheidungen: Regelungsbedarf im VwVG?, Die Einführung des Begriffs «automatisierte Einzelentscheidungen» und dessen Auswirkungen auf das Verwaltungsverfahrensrecht, Masterarbeit vom 12.03.2020 an der Universität Basel

**Ehrenzeller Bernhard/Schindler Benjamin/ Schweizer Rainer J./Vallender Klaus A.** (Hrsg.) (2014), Bundesverfassung St. Galler Kommentar, 3. Aufl., Zürich (zit. Bearbeiter/in, 2014, N … zu Art. … BV)

**Eifert Martin** (2006), Electronic Government: Das Recht der elektronischen Verwaltung, Baden-Baden

**Engelmann Jan/Puntschuh Michael** (2020), KI im Behördeneinsatz: Erfahrungen und Empfehlungen, Kompetenzzentrum Öffentliche IT (Hrsg), abrufbar unter: https://www.oeffentliche-it.de/documents/10181/14412/Best+Practices+beim+Einsatz+Künstlicher+Intelligenz+in+der+öffentlichen+Verwaltung

**Epiney Astrid** (2011), in: Belser Eva Maria/Epiney Astrid/ Waldmann Bernhard (Hrsg.), § 9 (Allgemeine Grundsätze)

**Ertel Wolfgang** (2016), Grundkurs Künstliche Intelligenz, Eine praxisorientierte Einführung, 4. Aufl., Wiesbaden

**Etscheid Jan** (2018), Automatisierungspotenziale in der Verwaltung, in: Mohabbat Kar Resa/Thapa Basanta/Paryecek Peter (Hrsg.), (Un)Berechenbar? Algorithmen und Automatisierung in Staat und Gesellschaft, Berlin, S. 126 ff.

**Etscheid Jan/von Lucke Jörn/Stroh Felix** (2020), Künstliche Intelligenz in der öffentlichen Verwaltung, Stuttgart

**Felzmann Heike et al.** (2019), Transparency You Can Trust: Transparency Requirements for Artificial Intelligence between Legal Norms and Contextual Concerns, in: Big Data & Society 6, no. 1, abrufbar unter: https://doi.org/10.1177/2053951719860542

**Fischer Claudio** (2020), Die digitale Steuerverwaltung, in: zsis) 2/2020, A6, S. 39 ff.

**Fischer Claudio/Daepp Annemarie** (2019), Digitalisierung am Beispiel der Steuerverwaltung des Kantons Bern, in: EF 4/19, S. 327 ff.

**Flick Uwe** (2004), Triangulation – Eine Einführung, Wiesbaden

**Floridi Luiano/Cowls Josh** (2019), A Unified Framework of Five Principles for AI in Society, abrufbar unter: https://doi.org/10.1162/99608f92.8cd550d1

**Flückiger Thomas** (2014), § 4 Verwaltungsverfahren, in: Steiger-Sackmann Sabine/Mosimann Hans-Jakob (Hrsg.), Recht der Sozialen Sicherheit, Sozialversicherungen, Opferhilfe, Sozialhilfe – Beraten und Prozessieren, Basel, S. 97 ff.

**Friedmann Batya/Nissenbaum Helen** (1996), Bias in Computer Systems, in: ACM Transactions of Information Systems, 03/1996, S. 330 ff.

**Gächter Thomas/Burch Stephanie** (2014), § 1 Nationale und internationale Rechtsquellen, in: Steiger-Sackmann, Sabine/Mosimann, Hans-Jakob (Hrsg.): Recht der Sozialen Sicherheit, Sozialversicherungen, Opferhilfe, Sozialhilfe – Beraten und Prozessieren, Basel, S. 3 ff.

**Gerstner Dominik** (2017), Predictive Policing als Instrument zur Prävention von Wohnungseinbruchdiebstahl: Evaluationsergebnisse zum Baden-Württembergischen Pilotprojekt P4, abrufbar unter: http://hdl.handle.net/11858/00-001M-0000-002E-384A-F

**Glaser Andreas** (2018), Einflüsse der Digitalisierung auf das schweizerische Verwaltungsrecht, in: SJZ 114/2018, S. 181 ff.

**Glaser Andreas/Ehrat Marco** (2019), E-Government-Gesetzgebung durch die Kantone – Integration in die Verfahrenskodifikation oder Auslagerung in Spezialerlasse?, in: LeGes 2019, S. 1 ff.

**Glass Philip** (2018), Gedanken zur Revision des DSG, abrufbar unter: https://www.datalaw.ch/gedanken-zur-revision-des-dsg/

**Goodman Bryce/Flaxman Seth** (2016), European Union regulations on algorithmic decision-making and a "right to explanation", Oxford, abrufbar unter: https://arxiv.org/pdf/1606.08813.pdf

**Görz Günther/Schmid Ute/Wachsmuth Ipke** (2014), Einleitung, in: Görz Günther/Schneeberger Josef/Schmid Ute (Hrsg.), Handbuch der Künstlichen Intelligenz, 5. Aufl., München, S. 1 ff.

**Griffel Alain** (2017), Allgemeines Verwaltungsrecht im Spiegel der Rechtsprechung, Zürich/Basel/Genf

**Griffel Alain** (Hrsg.) (2014), Kommentar zum Verwaltungsrechtspflegegesetz des Kantons Zürich (VRG), 3. Aufl., Zürich/Basel/Genf (zit. Bearbeiter/in, 2014, N … zu § … VRG)

**Guckelberger Annette** (2019), Öffentliche Verwaltung im Zeitalter der Digitalisierung, Baden-Baden

**Häfelin Ulrich/Müller Georg/Uhlmann Felix** (2020), Allgemeines Verwaltungsrecht, 8. Aufl., Zürich/St. Gallen

**Hagendorff Thilo** (2019), Maschinelles Lernen und Diskriminierungen: Probleme und Lösungsansätze, in: ÖZS 01/2019, S. 53 ff.

**Hammerschmid Gerhard/Raffer Christian** (2020), Künstliche Intelligenz im öffentlichen Sektor: Potenziale nutzen, Risiken bedenken, abrufbar unter: https://publicgovernance.de/media/KI_Oeffentliche_Verwaltung.pdf

**Hanania Pierre-Adrien/Knobloch Tobias** (2020), Künstliche Intelligenz im öffentlichen Sektor - Teil 1, abrufbar unter: https://www.capgemini.com/de-de/wp-content/uploads/sites/5/2020/10/PublicGoesAI-PoV-Part1-23122020.pdf

**Harasgama Rehana C.** (2017), Erfahren – Wissen – Vergessen, Zur zeitlichen Dimension des staatlichen Informationsanspruches, Zürich/St. Gallen

**Hoffmann Jens/Glaz-Ocik Justine** (2012), DyRiAS-Intimpartner: Konstruktion eines online gestützten Analyse-Instrument zur Risikoeinschätzung von tödlicher Gewalt gegen aktuelle oder frühere Intimpartnerinnen, in: Polizei und Wissenschaft, 2/2012, S. 45 ff.

**Jaag Tobias/Rüssli Markus** (2019), Staats- und Verwaltungsrecht des Kantons Zürich, 5. Aufl.

**Jobin Anna/Ienca Marcello/Vayena Effy** (2019), The Global Landscape of AI Ethics Guidelines, in: Nature Machine Intelligence 1, no 9, S. 389 ff., abrufbar unter: https://www.nature.com/articles/s42256-019-0088-2

**Karlen Peter** (2018), Schweizerisches Verwaltungsrecht – Gesamtdarstellung unter Einbezug des europäischen Kontextes, Zürich

**Kayser-Bril Nicolas** (2019), Austria's employment agency rolls out discriminatory algorithm, sees no problem, abrufbar unter: https://algorithmwatch.org/en/story/austrias-employment-agency-ams-rolls-out-discriminatory-algorithm/

**Kessler Rainer/Oberlin Jutta Sonja** (2020), Künstliche Intelligenz: Quo Vadis?, in: Compliance Berater 2020, S. 89 ff.

**Kiener Regina/Kälin Walter/Wyttenbach Judith** (2018), Grundrechte, 3. Aufl., Bern

**Kiener Regina/Rütsche Bernhard/Kuhn Mathias** (2015), Öffentliches Verfahrensrecht, 2. Aufl., Zürich/St. Gallen

**Kieser Ueli** (2019), Leistungen der Sozialversicherung, 3. Aufl., Zürich

**Kieser, Ueli** (2020), Kommentar zum Bundesgesetz über den Allgemeinen Teil des Sozialversicherungsrechts ATSG, 4. Aufl., Zürich (zit. Kieser, 2020, N … zu Art. … ATSG)

**Kirn Stefan/Müller-Hengstenberg Claus** (2013), Intelligente (Software-)Agenten: Von der Automatisierung zur Autonomie) Verselbständigung technischer Systeme, in: MMR 2013, S. 225 ff.

**Knobloch Tobias** (2018), Vor die Lage kommen: Predictive Policing in Deutschland – Chancen und Gefahren datenanalytischer Prognosetechnik und Empfehlungen für den Einsatz in der Polizeiarbeit, abrufbar unter: https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/predictive.policing.pdf

**Kolleck Alma/Orwat Carsten** (2020), Mögliche Diskriminierung durch algorithmische Entscheidungssysteme und maschinelles Lernen – ein Überblick, Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag, abrufbar unter: https://www.tab-beim-bundestag.de/de/pdf/publikationen/berichte/TAB-Hintergrundpapier-hp024.pdf

**Kroll Joshua A. et al.** (2016/2017), Accountable Algorithms, in: University of Pennsylvania Law Review 165, S. 633 ff., abrufbar unter: https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3/

**Krüger Jochen/Vogelgesang Stephanie/Adam Lena-Marie** (2020), Verantwortungsbewusste Digitalisierung, gerichtliche Entscheidungen und der Gedanke des fairen Verfahrens, in: Jusletter IT vom 28. Februar 2020

**Leese Matthias** (2018), Predictive Policing in der Schweiz: Chancen, Herausforderungen, Risiken, in: Bulletin zur schweizerischen Sicherheitspolitik 2018, S. 57 ff.

**Leslie David** (2019), Understanding Artificial Intelligence Ethics and Safety, A guide for the responsible design and implementation of AI systems in the public sector, The Alan Turing Institute, abrufbar unter: https://doi.org/10.5281/zenodo.3240529

**Locher Peter** (2015), Kommentar zum Bundesgesetz über die direkte Bundessteuer, III. Teil – Art. 102 – 222 DBG, Basel (zit. Locher, 2015, N … zu Art. … DBG)

**Loi Michele** (2015), Technological Unemployment and Human Disenhancement, in: Ethics and Information Technology, S. 1 ff., abrufbar unter: https://doi.org/10.1007/s10676-015-9375-8

**Loi Michele** (2020), People Analytics Must Benefit the People, An Ethical Analysis of Data-Driven Algorithmic Systems in Human Resources Management, abrufbar unter: https://algorithmwatch.org/wp-content/uploads/2020/03/AlgorithmWatch_AutoHR_Study_Ethics_Loi_2020.pdf

**Loi Michele/Ferrario Andrea/Viganò Eleonora** (2020), Transparency as Design Publicity: Explaining and Justifying Inscrutable Algorithms, in: Ethics and Information Technology, abrufbar unter: https://doi.org/10.1007/s10676-020-09564-w

**Loi Michele/Heitz Christoph/Christen Markus** (2020), A Comparative Assessment and Synthesis of Twenty Ethics Codes on AI and Big Data, in: 2020 7th Swiss Conference on Data Science (SDS), S. 41 ff., abrufbar unter: https://ieeexplore.ieee.org/document/9145014

**Mainzer Klaus** (2019), Künstliche Intelligenz – Wann übernehmen die Maschinen?, 2. Aufl., Berlin

**Martini Mario** (2019), Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, Berlin

**Martini Mario/Nink David** (2017), Wenn Maschinen entscheiden… – vollautomatisierte Verwaltungsverfahren und der Persönlichkeitsschutz, in: NVwZ – Extra 36/2017 Nr. 10, S. 1 ff.

**Meier-Mazzucato Giorgio** (2015), Steuern Schweiz, Grundriss zu den eidgenössischen und kantonalen Steuern mit Beispielen und Darstellungen, Bern

**Meyer Christian** (2019), Die Mitwirkungsmaxime im Verwaltungsverfahren des Bundes – Ein Beitrag zur Sachverhaltsfeststellung als arbeitsteiligem Prozess, Diss. Luzern

**Mittelstadt Brent/Russell Chris/Wachter Sandra** (2019), Explaining Explanations in AI, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, S. 279 ff., abrufbar unter: https://doi.org/10.1145/3287560.3287574

**Müller Jörg Paul/Schefer Markus** (2008), Grundrechte in der Schweiz, Im Rahmen der Bundesverfassung, der EMRK und der UNO-Pakte, 4. Aufl., Bern

**Nufer Marianne** (2019/2020), Künstliche Intelligenz in der Steuerveranlagung, in: ASA 88 (2019/2020), S. 259 ff.

**Opiela Nicole/Mohabbat Kar Resa/Thapa Basanta/Weber Mike** (2018), Exekutive KI 2030, Vier Zukunftsszenarien für Künstliche Intelligenz in der öffentlichen Verwaltung, Kompetenzzentrum Öffentliche IT (Hrsg.), abrufbar unter: https://www.oeffentliche-it.de/documents/10181/14412/Exekutive+KI+2030+-+Vier+Zukunftsszenarien+für+Künstliche+Intelligenz+in+der+öffentlichen+Verwaltung

**Ramge Thomas** (2018), Mensch und Maschine, Wie Künstliche Intelligenz und Roboter unser Leben verändern, 2. Aufl., Ditzingen

**Rawls John** (1999), A Theory of Justice, 2. Aufl., Cambridge

**Rechsteiner David** (2018), Der Algorithmus verfügt, Verfassungs- und verwaltungsrechtliche Aspekte automatisierter Einzelentscheidungen, in: Jusletter vom 26. November 2018

**Reich Markus** (2020), Steuerrecht, 3. Aufl., Zürich

**Reichwald Julian/Pfisterer Dennis** (2016), Autonomie und Intelligenz im Internet der Dinge, Möglichkeiten und Grenzen autonomer Handlungen, in: CR 3/2016, S. 208 ff.

**Reichwald Julian/Pfisterer Dennis** (2016), Autonomie und Intelligenz im Internet der Dinge, Möglichkeiten und Grenzen autonomer Handlungen, in: CR 2/2016, S. 208 ff.

**Reisman Dillon et al.** (2018), Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability, AI Now Institute, abrufbar unter: https://ainowinstitute.org/aiareport2018.pdf

**René Rhinow** (1983), Vom Ermessen im Verwaltungsrecht: eine Einladung zum Nach- und Umdenken Teil 2, in: recht 1983, S. 87 ff.

**Rhinow René/Koller Heinrich/Kiss Christina/Thurnherr Daniela/Brühl-Moser Denise** (2014), Öffentliches Prozessrecht, Grundlagen und Bundesrechtspflege, 3. Aufl., Basel

**Rhinow René/Schefer Markus/Uebersax Peter** (2016), Schweizerisches Verfassungsrecht, 3. Aufl., Basel

**Richner Felix/Frei Walter/Kaufmann Stefan/Meuter Hans Ulrich** (2013), Kommentar zum Zürcher Steuergesetz, 3. Aufl., Zürich (zit. Richner/Frei/Kaufmann/Meuter, N … zu Art. … StG)

**Ringeisen Peter/Bertolosi-Lehr Andrea/Demaj Labinot** (2018), Automatisierung und Digitalisierung in der öffentlichen Verwaltung: digitale Verwaltungsassistenten als neue Schnittstelle zwischen Bevölkerung und Gemeinwesen, in: YSAS 91, 2018, S. 51 ff., abrufbar unter: https://doi.org/10.5334/ssas.123

**Rosenthal David** (2020): Das neue Datenschutzgesetz, in: Jusletter vom 16. November 2020

**Rudin Beat** (2017), Anpassungsbedarf in den Kantonen, in: digma 2017, S. 58 ff.

**Sachs Michael/Schmitz Heribert** (Hrsg.) (2018), Kommentar zum Verwaltungsverfahrensgesetz: VwVfG, 9. Aufl., München (zit. Bearbeiter/in, 2018, N … zu § … VwVfG)

**Sägesser Thomas** (2014), Kommentierungen zu Art. 26 in: Graf Martin/Theler Cornelia/von Wyss Moritz (Hrsg.), Parlamentsrecht und Parlamentspraxis der Schweizerischen Bundesversammlung, Kommentar zum Parlamentsgesetz (ParlG) vom 13. Dezember 2002, Basel (zit. Sägesser, 2014, N … zu Art. 26 ParlG)

**Samek Wojciech/Montavon Grégoire/Vedaldi Andrea/Hansen Lars Kai/Müller Klaus-Robert** (Hrsg.) (2019), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Cham

**Scholta Hendrik/Mertens William/Kowalkiewicz Marek/Becker Jörg** (2019), From one-stop shop to no-stop shop: An e-government stage model, in: Government Information Quarterly 36, S. 11 ff.

**Schindler Benjamin** (2010), Verwaltungsermessen, Zürich/St. Gallen

**Simmler Monika/Brunner Simone/Schedler Kuno** (2020), Smart Criminal Justice – Eine empirische Studie zum Einsatz von Algorithmen in der Schweizer Polizeiarbeit und Strafrechtspflege, St. Gallen, abrufbar unter: https://www.alexandria.unisg.ch/261666/1/Simmler%20et%20al._Smart%20Criminal%20Justice_Forschungsbericht%20vom%2010.12.2020.pdf

**Simmler Monika** (Hrsg.) (2021), Smart Criminal Justice. Der Einsatz von Algorithmen in der Polizeiarbeit und in der Strafrechtspflege, Basel

**Söbbing Thomas** (2018), Künstliche Intelligenz im HR-Recruiting-Prozess: Rechtliche Rahmenbedingungen und Möglichkeiten, in: InTer 2018, S. 64 ff.

**Stiemerling Oliver** (2015), «Künstliche Intelligenz» – Automatisierung geistiger Arbeit, Big Data und das Internet der Dinge, Eine technische Perspektive, in: CR 12/2015, S. 762 ff.

**Thouvenin Florent/Braun Binder Nadja** (i. V.), Datenschutzerklärungen für den Bund

**Thurnherr Daniela** (2013), Verfahrensgrundrechte und Verwaltungshandeln, Habil. Basel

**Treuthardt Daniel/Loewe-Baur Mirjam/Kröger Melanie** (2018), Der Risikoorientierte Sanktionenvollzug (ROS) – aktuelle Entwicklungen, in: SKZ 2/2018, S. 24 ff.

**Tschannen Pierre/Zimmerli Ulrich/Müller Markus** (2014), Allgemeines Verwaltungsrecht, 4. Aufl., Bern

**Uhlmann Felix/Stojanovic Jasna** (2017), Vertrauen im Finanzmarktrecht aus öffentlich-rechtlicher Sicht, in: SZW 2017

**Vieth Kilian/Wagner Ben** (2017), Teilhabe, ausgerechnet, Wie algorithmische Prozesse Teilhabechancen beeinflussen können, Gütersloh

**Vokinger Kerstin Noëlle/Mühlematter Urs Jakob/Becker Anton/Boss Andreas/Reutter Mark A./Szucs Thomas D.** (2017), Artificial Intelligence und Machine Learning in der Medizin, in: Jusletter vom 28. August 2017

**Von Lucke Jörn** (2019), Disruptive Modernisierung von Staat und Verwaltung durch den gezielten Einsatz von smarten Objekten, cyberphysischen Systemen und künstlicher Intelligenz, Digitalisierung von Staat und Verwaltung - Gemeinsame Fachtagung Verwaltungsinformatik (FTVI) und Fachtagung Rechtsinformatik (FTRI) 2019, S. 49 ff.

**Von Lucke Jörn/Etscheid Jan** (2020), Wie Ansätze Künstlicher Intelligenz die öffentliche Verwaltung und die Justiz verändern könnten, in: Jusletter vom 21. Dezember 2020

**Waldmann Bernhard/Belser Eva Maria/Epiney Astrid** (Hrsg.) (2015), Basler Kommentar, Bundesverfassung (zit. Bearbeiter/in, 2015, N … zu Art. … BV)

**Waldmann Bernhard/Wiederkehr René** (2019), Allgemeines Verwaltungsrecht, Zürich/Basel/Genf

**Weber Rolf H.** (2019), Digitalisierung und der Kampf ums Recht, in: APARIUZ 2019, S. 1 ff.

**Weber Rolf H.** (2020), Automatisierte Entscheidungen: Perspektive Grundrechte, in: SZW 2020, S. 18 ff.

**Weber Rolf H./Henseler Simon** (2020), Regulierung von Algorithmen in der EU und in der Schweiz, in: EuZ 2020, S. 28 ff.

**Weber-Dürler Beatrice** (2001), Neuere Entwicklung des Vertrauensschutzes, in: ZBI 103/2001

**Widmer Dieter** (2019), Die Sozialversicherung in der Schweiz, 12. Aufl., Zürich

**Wiederkehr René** (2010), Die Begründungspflicht nach Art. 29 Abs. 2 BV und die Heilung bei Verletzung, in: ZBI 111/2010, S. 481 ff.

**Wiederkehr René** (2016), Öffentliches Verfahrensrecht, Bern

**Wischmeyer Thomas** (2018): Regulierung intelligenter Systeme, in: Di Fabio Udo/Eifert Martin/Huber Peter M. (Hrsg.), AöR 143/2018 Nr. 1, S. 1 ff.

**Zanella Andrea/Bui Nicola/Castellani Angelo/Vangelista Lorenzo/Zorzi Michele** (2014), Internet of Things for Smart Cities, in: IEEE, Vol. 1 No. 1, 2014, abrufbar unter: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6740844

**Zarsky Tal Z.** (2013), Transparent Predictions, in: University of Illinois Law Review, S. 1503 ff., abrufbar unter: https://illinoislawreview.org/wp-content/ilr-content/articles/2013/4/Zarsky.pdf

**Zweifel Martin** (2017), Kommentierung der Art. 109–119, 122–135 DBG, Zweifel Martin/Beusch Michael, Bundesgesetz über die direkte Bundessteuer (DBG), Kommentar zum Schweizerischen Steuerrecht, 3. Aufl., Basel 2017 (zit. Zweifel, 2017, N … zu Art. … DBG)

**Zweifel Martin/Casanova Hugo/Beusch Michael/Hunziker Silvia** (2018), Schweizerisches Steuerverfahrensrecht – Direkte Steuern, 2. Aufl., Zürich/Basel/Genf

**Zweifel Martin/Hunziker Silvia** (2008/2009), Steuerverfahrensrecht, Beweislast, Drittvergleich, «dealing at arm's length», Art. 29 Abs. 2 BV, Art. 58 DBG, Beweis und Beweislast im Steuerverfahren bei der Prüfung von Leistung und Gegenleistung unter dem Gesichtswinkel des Drittvergleichs («dealing at arm's length»), in: ASA 77 (2008/09), S. 657 ff.

**Zweig Katharina** (2016), 1. Arbeitspapier, Was ist ein Algorithmus?. Berlin, abrufbar unter: https://algorithmwatch.org/publication/arbeitspapier-was-ist-ein-algorithmus/

**Zweig Katharina** (2019a), Algorithmische Entscheidungen: Transparenz und Kontrolle, in: Analyse und Argumente, Digitale Gesellschaft, Nr. 338

**Zweig Katharina** (2019b), Ein Algorithmus hat kein Taktgefühl, Wo Künstliche Intelligenz sich irrt, warum uns das betrifft und was wir dagegen tun können, München

**Zavadil Andreas** (2020), Datenschutzrechtliche Zulässigkeit des „AMS-Algorithmus", abrufbar unter: https://www.dsb.gv.at/dam/jcr:dcc945ef-8666-4859-9362-060dedb12f1c/Newsletter-4-2020.pdf Kapitel 6

## / Materialienverzeichnis

**AI Now Institute et al.,** Using Procurement Instruments to Ensure Trustworthy, abrufbar unter: https://assets.mofoprod.net/network/documents/Using_procurement_instruments_to_ensure_trustworthy_AI.pdf (zit. AI Now Institute) (zit. AI Now Institute et al.)

**AlgorithmWatch** (2020), Automating Society Report 2020, A report by AlgorithmWatch in cooperation with Bertelsmann Stiftung, supported by the Open Society Foundations, Berlin, abrufbar unter: https://automatingsociety.algorithmwatch.org (zit. Automating Society Report 2020)

**Amtsblatt Zürich** (2004), Antrag des Regierungsrates vom 21. Juli 2004 zur Änderung des Steuergesetzes, Amtsblatt Kanton Zürich, S. 810 ff. (zit. ABl ZH 2004)

**Automated Decision Systems Task Force** (2019), New York City Automated Decision Systems Task Force Report, abrufbar unter: https://www1.nyc.gov/assets/adstaskforce/downloads/pdf/ADS-Report-11192019.pdf (zit. New York City, 2019)

**Bitkom/DFKI** (2017), Deutsches Forschungszentrum für Künstliche Intelligenz (Hrsg.)

**Entscheidungsunterstützung mit Künstlicher Intelligenz,** abrufbar unter: https://www.bitkom.org/sites/default/files/file/import/171012-KI-Gipfelpapier-online.pdf (zit. Bitkom/DFKI, 2017)

**Botschaft vom 15. September 2017** zum Bundesgesetz über die Totalrevision des Bundesgesetzes über den Datenschutz und die Änderung weiterer Erlasse zum Datenschutz, BBl 2017 6941 ff. (zit. Botschaft E-DSG)

**Botschaft vom 19. Februar 2003 zur Änderung des Bundesgesetzes über den Datenschutz** (DSG) und zum Bundesbeschluss betreffend den Beitritt der Schweiz zum Zusatzprotokoll vom 8. November 2021 zum Übereinkommen zum Schutz des Menschen bei der automatischen Verarbeitung personenbezogener Daten bezüglich Aufsichtsbehörden und grenzüberschreitende Datenübermittlung, BBl 2003 2101 ff. (zit. Botschaft DSG und Zusatzprotokoll)

**Botschaft vom 24. September 1965** des Bundesrates an die Bundesversammlung über das Verwaltungsverfahren, BBl 1965 1348 ff. (zit. Botschaft VwVG)

**Botschaft vom 25. Mai 1983** zu den Bundesgesetzen über die Harmonisierung der direkten Steuern der Kantone und Gemeinden sowie über die direkte Bundessteuer, BBl 1983 III 1 ff. (zit. Botschaft Steuerharmonisierung)

**CAHAI** (Ad hoc Committee on Artificial Intelligence) (2020), Towards Regulation of AI Systems, Global perspectives on the development of a legal framework on Arificial Intelligence (AI) systems based on the Council of Europe's standards on human rights, democracy and the rule of law, abrufbar unter: https://rm.coe.int/prems-107320-gbr-2018-compli-cahai-couv-texte-a4-bat-web/1680a0c17a (zit. CAHAI 2020)

**Cities for Digital Rights** (2020), Declaration of Cities Coalition for Digital Rights, abrufbar unter: https://citiesfordigitalrights.org/declaration (zit. Cities for Digital Rights, 2020)

**Council of Europe** (2020), European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment, Strasbourg 2020, abrufbar unter: https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c (zit. Council of Europe, 2020, Ethical Charter AI)

**Council of Europe** (2020), Recommendation CM/Rec(2020)1 of the Committee of Ministers to Member States on the Human Rights Impacts of Algorithmic Systems, Strasbourg 2020, abrufbar unter: https://rm.coe.int/09000016809e1154 (zit. Council of Europe, CM/Rec (2020)1)

**Dataethical Thinkdotank** (2021), White Paper: Data Ethics in Public Procurement, abrufbar unter: https://dataethics.eu/publicprocurement/ (zit. Dataethical Thinkdotank, 2021)

**Datenethikkommission** (2019), Gutachten der Datenethikkommission der Bundesregierung, Berlin, abrufbar unter: https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf;jsessionid=C69B0F3A9EEC177FFB7771A68F72E30B.1_cid364?__blob=publicationFile&v=6

**Government Digital Service and Office for Artificial Intelligence UK** (2019), A Guide to Using Artificial Intelligence in the Public Sector/ Understanding Artificial Intelligence Ethics and Safety, abrufbar unter: https://www.gov.uk/guidance/understanding-artificial-intelligence-ethics-and-safety. (Government UK, 2019)

**Government of Canada and Treasury Board Secretariat** (2019), Directive on Automated Decision Making, abrufbar unter: https://assets.mofoprod.net/network/documents/Using_procurement_instruments_to_ensure_trustworthy_AI.pdf (zit. Government of Canada, 2019)

**Government of New Zealand** (2020), Algorithm Charter for Aotearoa New Zealand, abrufbar unter: https://data.govt.nz/use-data/data-ethics/government-algorithm-transparency-and-accountability/algorithm-charter (zit. Government of New Zealand, 2020)

**Independent High-Level Expert Group On Artificial Intelligence Set Up By The European Commission** (2019), Ethics Guidelines for Trustworthy AI" (European Commission - Digital Single Market), abrufbar unter: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (zit. Ethics Guidelines for Trustworthy AI)

**OECD** (2019), Artificial Intelligence in Society, OECD Publishing, Paris, abrufbar unter: https://www.oecd-ilibrary.org/docserver/eedfee77-en.pdf?expires=1610099411&id=id&accname=ocid195445&checksum=5D406E9B3F37E2B2CA6328E248A2F9AE (zit. OECD, 2019)

**SBFI** (2019), Bericht der interdepartementalen Arbeitsgruppe «Künstliche Intelligenz» an den Bundesrat über die Herausforderungen der Künstlichen Intelligenz (zit. Bericht Herausforderungen 2019)

**Schweizerische Eidgenossenschaft** (2020), der Bundesrat, Leitlinien „Künstliche Intelligenz" für den Bund, Orientierungsrahmen für den Umgang mit künstlicher Intelligenz in der Bundesverwaltung, vom 25. November 2020, abrufbar unter: https://www.admin.ch/gov/de/start/dokumentation/medienmitteilungen.msg-id-81319.html (zit. Bundesrat Leitlinien KI 2020)

**Secrétariat du Grand Conseil,** Projet présenté par le Conseil d'Etat Date de dépôt: 29 août 2018 , PL 12386 Projet de loi, «https://ge.ch/grandconseil/data/texte/PL12386.pdf» (besucht am: 1. Januar 2021) (zit. PL 12386 Projet de loi)

**TA-SWISS KI** (2020), Wenn Algorithmen für uns Entscheiden: Chancen und Risiken der Künstlichen Intelligenz, TA-SWISS Publikationsreihe (Hrsg.): TA 72/2020, Zürich (zit. TA-SWISS KI, 2020)

**World Economic Forum** (2020), AI Government Procurement Guidelines, abrufbar unter: http://www3.weforum.org/docs/WEF_AI_Procurement_in_a_Box_AI_Government_Procurement_Guidelines_2020.pdf (zit. WEF, 2020)

# Nachwort

Diese Studie ist im Zeitraum August 2020 bis Februar 2021 entstanden – ein für wissenschaftliche Projekte eher kurzer Zeitraum. Erschwerend kam hinzu, dass angesichts der Reisebeschränkungen und Restriktionen hinsichtlich physischer Treffen die ursprünglich als Präsenzveranstaltungen geplanten Workshops und Interviews ausschließlich online durchgeführt werden konnten. Ohne ein hoch motiviertes Team wäre diese Studie nicht möglich gewesen. Ein immenses Dankeschön gebührt deshalb den Mitwirkenden aus dem Team der juristischen Fakultät der Universität Basel, namentlich Catherine Egli, Laurent Freiburghaus, Eliane Kunz, Nina Laukenmann und Liliane Obrecht sowie den Beteiligten seitens AlgorithmWatch, namentlich Jessica Wulf für die Konzeption, Durchführung und Auswertung der Interviews und Dr. Michele Loi, der als Senior Research Advisor bei AlgorithmWatch Schweiz gemeinsam mit Dr. Anna Mätzener, Leiterin von AlgorihmWatch Schweiz, für die ethischen Erwägungen zuständig war.

Das Interesse des Kantons Zürich an der Thematik manifestiert sich einerseits in der Tatsache, dass diese Studie in Auftrag gegeben und namentlich durch Lukas Weibel, den Projektleiter in der Staatskanzlei des Kantons Zürich, stets kompetent und zielführend begleitet wurde – wofür wir ihm zu großem Dank verpflichtet sind. Andererseits konnten wir auf die Unterstützung verschiedener Expertinnen und Experten aus dem Kanton und der Stadt Zürich zählen. Wir danken Peter Seidler (Steuerverwaltung, Kanton Zürich), Michael Boller (Amt für Raumentwicklung, Kanton Zürich), Michael Bächinger (Sozialversicherungsanstalt, Kanton Zürich) und Dr. Dominika Blonski (Datenschutzbeauftragte, Kanton Zürich) sowie Rolf Brühlmann (Kompetenzzentrum für Organisation und Information, Stadt Zürich) für ihre Bereitschaft, uns als Interviewpartner zur Verfügung zu stehen. Ein besonderer Dank gilt außerdem Felix Bührer (Compliancebeauftragter, Kanton Zürich), Sandra Vogel (Juristische Mitarbeiterin Datenschutzbeauftragte, Kanton Zürich), Stefan Langenauer, Matthias Mazenauer und Dr. Christian Ruiz (alle Statistisches Amt der Justizdirektion, Kanton Zürich) sowie Christian Häberli (AWK Group AG), die uns im Rahmen eines Workshops Feedback zu den beiden Ethik-Checklisten gegeben haben. Wir danken schließlich den verschiedenen Personen in kantonalen Steuerverwaltungen und einzelnen Unternehmen, die wir im Laufe der Untersuchung telefonisch kontaktiert haben und die uns bereitwillig für Hintergrundgespräche zur Verfügung standen.

Nadja Braun Binder, Matthias Spielkamp

Automatisierte Entscheidungssysteme im öffentlichen Sektor
Ein Impact-Assessment-Tool für die öffentliche Verwaltung

Von Michele Loi, in Zusammenarbeit mit Anna Mätzener, Angela Müller und Matthias Spielkamp

September 2022

Online verfügbar unter https://algorithmwatch.org/de/impact-assessment-adm-oeffentliche-verwaltung

Herausgeber:
AW AlgorithmWatch gGmbH
Linienstr. 13
10178 Berlin
https://algorithmwatch.org
info@algorithmwatch.org

AlgorithmWatch
Spindelstr. 2
8041 Zürich
https://algorithmwatch.ch
info@algorithmwatch.ch

Layout:
Beate Autering, Beate Stangl,
www.beworx.de