ENDBERICHT

AI-Act for Austria (A3) Studie zur Umsetzung des AI-Acts in kritischen Infrastrukturen Österreichs

Finanziert im Sicherheitsforschungs-Förderprogramms KIRAS des Bundesministeriums für Finanzen.

Inhaltsverzeichnis

Einleitung 4
Rechtliche Analyse 5
Risikomanagement bei KI 5
Einleitung 5
Forschungsansatz und Ergebnisse 7
State-of-the-Art von AI Risikomanagement 8
AI Product Lifecycle 10
ITIL v4 and KI-Systeme 10
AI Product Lifecycle: Schritte 11
AI Product Lifecycle mit Beispiel 14
Holistischer KI-Risikomanagement-Ansatz 14
Data Science Trajectories Framework 18
Einführung 18
Kategorien 18
Explorative Aktivitäten 18
Zielgerichtete Aktivitäten 20
Datenmanagementaktivitäten 21
Komponenten 22
Explorative Komponenten 22
Zielgerichtete Komponenten (CRISP-DM) 57
Komponenten der Datenverwaltung 94
Risiko-Register 98
Risiken für das KI-Modell und Daten 98
M-01 - Model Inversion 98
M-02 - Poisoning Angriffe 99
M-03 - Model Theft/Extraction 102
M-04 - Model Skewing 104
M-05 - Input Manipulation 106
M-06 - Prompt Injection 108
M-07 - Datendrift 110
M-08 - Concept Drift 112
M-09 - Mangel an gekennzeichneten (Trainings)daten 114
M-10 - Overfitting 116
M-11 - Verzerrung bei der Merkmalsauswahl 118
Regulatorische und wirtschaftliche Herausforderungen 120

B-01 - Datenschutz 120
B-02 - Technologieabhängigkeit 122
Ethische und gesellschaftliche Herausforderungen 124
G-01 - Diskriminierung und Bias 124
G-02 - Transparenz und Erklärbarkeit 126
G-03 - Manipulation, Desinformation und Social Engineering 128
G-04 - Falsche Erwartungen an KI 130
Explorative Szenarioanalyse – Methodische Aspekte 132
Grundlagen und Basisansatz nach Geschka 133
Methodische Kritik 136
Finale Methodik der ESA nach A3 139
Explorative Szenarioanalyse – Durchführung & Ergebnisse 144
Ablauf der Durchführung der ESA 144
Interview-Phase 145
Auswahl der Expert*innen 146
Fragenkatalog und Ablauf der Interviews 147
Auswertungsphase 150
Isolierte Faktoren und Faktorausprägungen 151
Generierte Szenarien 160
Szenario 1 – „Mittelweg“ 161
Szenario 2 – „Abstieg“ 164
Szenario 3 – „Passiv Aktiv“ 167
Szenario 4 – „Wild AI West“ 170
Szenario 5 – „Geopolitische Instabilität“ 173
Beschaffungsleitfaden 179
Einleitung 179
Grundlagen und Rahmenbedingungen 179
Rechtliche und regulatorische Grundlagen 179
Organisatorische und technische Voraussetzungen 180
Ermittlung der Anforderungen 180
Make-Buy-or Transfer Guide für KI 181
Einleitung 181
Make-Buy-or Transfer Guide für KI 182
Option 1: Make 184
Option 2: Buy 185
Option 3: Transfer 185

Fazit 186
Handlungsempfehlungen 187
Option 1: Make 190
Option 2: Buy 191
Option 3: Transfer 191
Prozessübersicht 193
Checkliste 193
Kriterien für die Entscheidung – Scoring-Modell 196
Logischer Aufbau des Scoring-Modells 196
Zusammenfassung und Nutzen der erweiterten Checklisten und Scoring-Modelle 197
Zusammenfassung 199
Anhang A – Leere Checkliste 200
Anhang B - Risiko Akteure 207
Anhang C – Anwendungsfälle von KI und ihre Risiken 210
Use Case 1: Analyse sozialer Netzwerke 210

Einleitung

Die rasante Entwicklung und zunehmende Integration von Künstlicher Intelligenz (KI) in verschiedenste gesellschaftliche, wirtschaftliche und politische Bereiche stellen Organisationen, Behörden und Unternehmen vor tiefgreifende Herausforderungen. Einerseits bietet die Implementierung von KI-Technologien enorme Chancen, Effizienzsteigerungen, verbesserte Entscheidungsfindung sowie erhöhte Sicherheit in zahlreichen Anwendungsfeldern. Andererseits gehen mit dem Einsatz von KI jedoch auch erhebliche Risiken einher, die insbesondere in kritischen Infrastrukturen und Hochrisiko-Anwendungen sorgfältig identifiziert, bewertet und behandelt werden müssen.

Das vorliegende Projekt, namens ***AI-Act for Austria (A3)*** setzt sich umfassend mit den Anforderungen und Herausforderungen auseinander, die durch den AI-Act der Europäischen Union entstehen. Der AI-Act bildet eine grundlegende regulatorische Basis, um den Einsatz von KI-Technologien zu steuern und die damit verbundenen Risiken effektiv zu minimieren. Insbesondere Hochrisiko-KI-Systeme unterliegen detaillierten und strengen Vorschriften, welche Organisationen vor erhebliche organisatorische und technische Herausforderungen stellen. Im Fokus stehen hierbei kritische Infrastrukturen (KRITIS), die aufgrund ihrer Bedeutung für die nationale Sicherheit und öffentliche Versorgung besonders hohe Anforderungen an Risikomanagement und Compliance erfüllen müssen.

Ziele des A3 Projekts:

(1) **Analyse der neu entstehenden EU-weiten Regularien** in Bezug auf die Nutzung von AI, speziell im Hochrisikobereich, bzw. im Anwendungsgebiet kritischer Infrastrukturen.

(2) **Empfehlungen** zu **organisatorischen, technischen und rechtlichen Maßnahmen**, die bei der Entwicklung und beim Einsatz von AI-Systemen über deren gesamten Lebenszyklus hinweg gewährleistet werden müssen

(3) **Herausforderungen bei Umsetzung** und Überprüfung der Anforderungen aus dem AI-Act in real existierenden kritischen Infrastrukturen **aufzuzeigen**

(4) Konkret **anwendbare Vorschläge** zu liefern

(5) Nutzung eines spezifisch adaptierten Ansatzes zur **explorativen Szenarien Analyse**

(6) Frage- und **Check-Listen**, **Beschaffungsleitfadens für sichere AI**

Im Rahmen dieses Projekts wurde umfassend untersucht, wie Betreiber kritischer Infrastrukturen und KMUs diesen Herausforderungen begegnen können. Zielsetzung war die Entwicklung konkreter Handlungsempfehlungen und Leitlinien, welche Organisationen dabei unterstützen, regulatorische Anforderungen strukturiert und effektiv umzusetzen. Die Kombination von juristischem und technischem Fachwissen sowie eine interdisziplinäre Herangehensweise ermöglichte eine fundierte Analyse und praxisorientierte Lösungen.

Der Bericht gliedert sich entsprechend in zwei Abschnitte:

- Dieser Berichtsteil der sich mit den technischen Aspekten, aber auch Aspekten des Risikomanagements beschäftigt, gliedert sich folgendermaßen auf:
- Der zweite Berichtsteil zu den rechtlichen Analysen ist aufgrund einer bewusst anders gewählten spezifischen Struktur als eigenes Dokument ausgegliedert (siehe auch das folgende Kapitel).

Rechtliche Analyse

Die rechtliche Analyse findet sich im Dokument „A3_Endbericht_Rechtlicher_Teil.pdf“. Sie wurde extern ausgelagert, um eine bessere Lesbarkeit des Gesamtdokuments zu gewährleisten.

Risikomanagement bei KI

Einleitung

Die wachsende Integration von Künstlicher Intelligenz (KI) in nahezu alle Lebens- und Arbeitsbereiche, insbesondere in kritischen Infrastrukturen wie Energieversorgung, Gesundheitswesen, Strafverfolgung, Verkehrsmanagement und Finanzsystemen, birgt neben großen Potenzialen auch signifikante Risiken. Ein umfassendes Risikomanagement ist notwendig, um sicherzustellen, dass der Einsatz dieser leistungsstarken Technologien ethisch vertretbar, gesellschaftlich akzeptabel und rechtlich konform erfolgt.

Regulatorische Compliance

Mit dem Inkrafttreten des AI-Acts der Europäischen Union werden Unternehmen und Behörden verpflichtet, spezifische regulatorische Anforderungen für Hochrisiko-KI-Systeme umzusetzen. Das Risikomanagement ist hierbei eine zentrale Komponente, welche die Einhaltung dieser Vorschriften gewährleistet. Der AI-Act fordert ausdrücklich ein strukturiertes Risikomanagementsystem (RMS), das den gesamten Lebenszyklus von KI-Systemen begleitet und dabei technische, organisatorische und rechtliche Aspekte berücksichtigt.

Ohne ein effektives Risikomanagement besteht die Gefahr erheblicher rechtlicher Konsequenzen wie Geldstrafen, rechtlicher Haftung und Reputationsverlust, insbesondere wenn die eingesetzten KI-Systeme nicht transparent, sicher und gerecht gestaltet sind.

Schutz der nationalen und öffentlichen Sicherheit

KI-Systeme, insbesondere solche in kritischen Infrastrukturen, haben das Potenzial, direkt Einfluss auf die öffentliche und nationale Sicherheit zu nehmen. Risikomanagement hilft dabei, potenzielle Bedrohungen frühzeitig zu identifizieren und proaktiv zu steuern, um Ausfälle, Fehlfunktionen oder Manipulationen zu verhindern.

Zum Beispiel könnten Cyberangriffe auf KI-gesteuerte Energieversorgungssysteme oder medizinische Einrichtungen katastrophale Folgen haben. Durch ein robustes Risikomanagement werden solche Risiken systematisch erfasst und Maßnahmen entwickelt, um Systeme widerstandsfähiger zu machen.

Minimierung von Verzerrungen (Bias) und Diskriminierung

KI-Modelle können unbeabsichtigte Vorurteile übernehmen und dadurch diskriminierende Entscheidungen treffen, welche gesellschaftliche Ungleichheiten verstärken könnten. Ein systematisches Risikomanagement hilft dabei, solche algorithmischen Verzerrungen frühzeitig zu erkennen, zu bewerten und gezielt zu reduzieren.

Ein Beispiel dafür ist die Verwendung von KI im Recruiting. Ohne angemessenes Risikomanagement könnten bestehende soziale Vorurteile in Einstellungsentscheidungen verstärkt werden, was zu ethischen und rechtlichen Herausforderungen führen könnte.

Sicherstellung der Fairness und Gerechtigkeit

KI-Systeme müssen transparent und nachvollziehbar sein, um Fairness und Vertrauen in Entscheidungsprozesse zu gewährleisten. Risikomanagement stellt sicher, dass Systeme überprüfbar und verständlich bleiben und unterstützt so das Vertrauen der Nutzer sowie der Öffentlichkeit.

Transparenz in KI-Entscheidungen verhindert das Risiko einer gesellschaftlichen Ablehnung und reduziert potentielle Konflikte zwischen Nutzern und Betreibern von KI-Systemen.

Widerstandsfähigkeit und Resilienz

Risikomanagement fördert die Entwicklung von robusten KI-Systemen, die gegenüber unerwarteten Ereignissen, Ausfällen oder Angriffen widerstandsfähig sind. Durch proaktive Maßnahmen und kontinuierliches Monitoring sind Organisationen besser vorbereitet, um auf Krisen und Vorfälle schnell zu reagieren und den Betrieb sicherzustellen.

Praxisbeispiele

Ein häufiges Beispiel für erfolgreiches Risikomanagement ist die Implementierung eines KI-basierten Systems zur Identifikation von Hassrede im Internet. Durch ein strukturiertes Risikomanagement können Risiken wie Fehlklassifikationen, Datenschutzverstöße oder ethische Probleme proaktiv behandelt werden.

Ein weiteres Beispiel ist der Einsatz von KI in autonomen Fahrzeugen, bei denen kontinuierliches Risikomanagement notwendig ist, um Risiken wie Unfälle oder Fehlentscheidungen zu minimieren.

Nutzen und Mehrwert von Risikomanagement

Die Implementierung eines systematischen Risikomanagements bietet Organisationen erhebliche Vorteile:

- **Rechtliche Sicherheit:** Einhaltung regulatorischer Anforderungen, Vermeidung von Strafzahlungen.
- **Schutz vor Schäden:** Minimierung finanzieller und operativer Schäden.
- **Vertrauen:** Erhöhung des Vertrauens bei Nutzern und Stakeholdern.
- **Innovationsfähigkeit:** Sichere und verantwortungsvolle Einführung neuer Technologien.

Das Risikomanagement für KI-Systeme ist nicht nur regulatorisch verpflichtend, sondern auch operativ und ethisch unerlässlich. Es schützt Organisationen und Gesellschaft gleichermaßen vor unerwünschten Folgen der KI-Technologie und stellt sicher, dass diese in verantwortungsvoller und nachhaltiger Weise genutzt werden kann.

Forschungsansatz und Ergebnisse

KI-Systeme zeichnen sich durch eine hohe Komplexität aus. Entscheidungen der Systeme sind oft schwer nachvollziehbar („Black Box Problem“). Zudem gibt es Herausforderungen hinsichtlich ethischer Fragen, Datenschutz, Transparenz und technischer Sicherheit.

Das Risikomanagement muss deshalb Methoden integrieren, die auf spezifische Herausforderungen der KI reagieren, wie etwa die Erklärbarkeit von Modellen (Explainable AI), proaktives Bias-Monitoring und kontinuierliche Audits der Systemleistung.

In Rahmen des A3 Projekts bezüglich Risikomanagement wurden die folgende 3 zentrale Instrumente erstellt, bzw. verbunden:

(1) AI Product Lifecycle (siehe Kapitel **AI Product Lifecycle**– Seite 10):
(2) DST Framework (siehe Kapitel **Data Science Trajectories Framework** – Seite 18)
(3) Risikomanagement durch das Risikoregister (siehe Kapitel **Risiko-Register** – Seite 98)

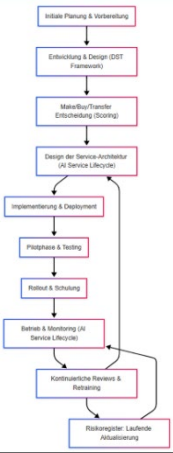
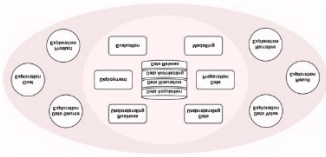
Projekt ergebnisse:	**AI Product Lifecycle**	**DST Framework**	**Risikomanagement durch das Risikoregister**
Beschreibung	Ein strukturierter, iterativer Prozess von der Bedarfserhebung bis zum Retiring eines KI-Systems, der sowohl technische als auch organisatorische und regulatorische Aspekte abdeckt.	Ergänzt diesen Lifecycle durch eine klare Unterteilung in explorative, zielorientierte und datenbezogene Aktivitäten	Das Risikoregister ist das zentrale Werkzeug, um entlang der gesamten Projekt- und Betriebsphasen potenzielle Gefahren zu identifizieren, zu bewerten und zu steuern.
Abbildung			
Zusammenhänge	**Der AI Service Lifecycle, das DST-Framework und das Risikoregister sind drei zentrale Instrumente, die zusammen ein umfassendes Management für KI-Projekte ermöglichen, jedoch unterschiedliche Schwerpunkte setzen.**		

Tabelle 1: Übersicht der Ergebnisse in Risikomanagement

Durch diese enge Verzahnung von AI Product Lifecycle, DST-Framework und Risikomanagement wird gewährleistet, dass KI-Systeme nicht nur effizient und leistungsfähig, sondern auch sicher, erklärbar und regelkonform entwickelt und betrieben werden – insbesondere in sensiblen, hochriskanten Bereichen mit staatlicher Beteiligung oder kritischer Infrastruktur.

State-of-the-Art von AI Risikomanagement

Der Einsatz von Künstlicher Intelligenz (KI) in kritischen Infrastrukturen bietet enormes Potenzial zur Steigerung der Effizienz und Sicherheit. Gleichzeitig entstehen spezifische Risiken, die sorgfältig verwaltet werden müssen, um negative Auswirkungen auf nationale Sicherheit, Datenschutz und ethische Standards zu verhindern. Im Kontext des AI-Acts der Europäischen Union stellt Risikomanagement eine zentrale Anforderung dar, insbesondere für sogenannte Hochrisiko-KI-Systeme.

Gemäß Artikel 9 des AI-Acts müssen Betreiber von Hochrisiko-KI-Systemen ein umfassendes Risikomanagementsystem (RMS) implementieren, das sämtliche Phasen des Lebenszyklus von KI-Systemen umfasst. Diese Verpflichtung dient dazu, Risiken, die sich aus der Verwendung von KI ergeben, zu minimieren und transparent zu steuern. Dabei sind die Betreiber gefordert, nicht nur technische, sondern auch organisatorische und rechtliche Aspekte im RMS zu integrieren.

Zur Unterstützung dieser Verpflichtung existieren verschiedene internationale Normen und Standards, die umfassende Methoden und Leitlinien für die Gestaltung und Umsetzung von Risikomanagementprozessen bieten. Zu den wichtigsten zählen:

- **ISO 31000 (Risk Management System):** Diese Norm stellt allgemeine Richtlinien für die Einrichtung und kontinuierliche Verbesserung eines Risikomanagementsystems bereit. Sie definiert Prozesse zur Risikoidentifikation, Risikoanalyse, Risikobewertung und Risikobehandlung und bietet einen strukturierten Rahmen für eine systematische Herangehensweise an Risiken.

- **ISO 23894 (AI Guidance on Risk Management):** ISO 23894 ergänzt allgemeine Risikomanagement-Standards gezielt für KI-Systeme. Sie bietet spezielle Leitlinien zur Identifikation, Analyse und Behandlung von KI-spezifischen Risiken wie Datensicherheit, Modellrobustheit und ethische Risiken. Diese Norm hilft insbesondere bei der Operationalisierung der Anforderungen aus dem AI-Act.

- **NIST AI Risk Management Framework:** Dieses Framework bietet einen integrierten Ansatz zur Bewertung und Steuerung von Risiken bei der Entwicklung und Nutzung von KI-Systemen. NIST legt besonderen Wert auf Transparenz, Verantwortung und Fairness der eingesetzten KI-Technologien und fördert einen ganzheitlichen Ansatz zur Risikobetrachtung, einschließlich der Betrachtung möglicher gesellschaftlicher Auswirkungen.

- **ISO 27005 (Information Security Risk Management):** Diese Norm fokussiert speziell auf Risiken im Bereich der Informationssicherheit und ergänzt somit das allgemeine Risikomanagement um eine essenzielle IT-sicherheitsorientierte Komponente. Sie stellt sicher, dass Betreiber von Hochrisiko-KI-Systemen systematisch Risiken im Kontext von Informations- und Datensicherheit erfassen und Maßnahmen zu deren Absicherung entwickeln.

Zusammen ermöglichen diese Standards eine ganzheitliche, systematische und transparente Steuerung von KI-Risiken, wodurch Compliance mit gesetzlichen Anforderungen, insbesondere des AI-Acts, gewährleistet wird.

Ein wirksames Risikomanagement umfasst mehrere Phasen:

1. Risikoidentifikation: Risiken werden systematisch erfasst und dokumentiert.
2. Risikoanalyse: Bewertung der Eintrittswahrscheinlichkeit und der möglichen Schadenshöhe.
3. Risikobehandlung: Entwicklung geeigneter Maßnahmen zur Reduzierung oder Eliminierung von Risiken.
4. Risikokontrolle: Kontinuierliche Überwachung der Risiken und regelmäßige Anpassung der Maßnahmen.
5. Diese Prozesse ermöglichen eine strukturierte und nachvollziehbare Herangehensweise an das Management von KI-Risiken.

Risiken im KI-Kontext

Die wesentlichen Risiken beim Einsatz von KI lassen sich in folgende Kategorien unterteilen:

- **Datengenerierte Risiken:**
  - Qualität der Daten, Vorurteile (Bias), Datenschutzprobleme
  - Beispiel: Unabsichtliche Veröffentlichung privater Informationen durch KI-Systeme
- **Technische Risiken:**
  - Modellinversion (extrahierte Informationen aus KI-Modellen)
  - Robustheit und Reproduzierbarkeit der Modelle
  - Transparenz und Erklärbarkeit der Entscheidungsprozesse
- **Operative Risiken:**
  - Fehlfunktionen und mangelnde Expertise im Umgang mit KI
  - Übermäßiges Vertrauen („Overreliance“) auf KI-Systeme
- **Gesellschaftliche Risiken:**
  - Manipulation und Fehlinformation durch automatisierte Prozesse
  - Ethikverstöße und Diskriminierungspotential (Bias)

Das Risikoregister dokumentiert systematisch Risiken, beschreibt Indikatoren, definiert Szenarien und schlägt Maßnahmen vor. Dieses Instrument dient als zentrale Datenbank, die kontinuierlich gepflegt und aktualisiert wird (siehe Kapitel Risikoregister unten).

AI Product Lifecycle

ITIL v4 and KI-Systeme

Im Rahmen des A3-Projektes wurde eine zentrale Aufgabe darin gesehen, ein robustes und umfassendes Lifecycle-Modell für KI-basierte Services und Produkte zu entwickeln. Die Entscheidung, hierfür die Logik des etablierten ITIL v4 Service Lifecycle-Modells zu verwenden, wurde auf Grundlage mehrerer wichtiger Faktoren getroffen, die in der Praxis von entscheidender Bedeutung sind. Das ITIL v4 Modell bietet eine strukturierte, bewährte und agile Herangehensweise an Service-Management, welche für den Bereich der Künstlichen Intelligenz (KI) besonders relevant erscheint.

Der ITIL v4 Lifecycle ist Teil des Service Value Systems (SVS) und fokussiert auf eine ganzheitliche Betrachtung, welche sowohl auf Flexibilität als auch auf Agilität ausgelegt ist. Der Lebenszyklus umfasst dabei fünf Hauptphasen: Service Strategy, Service Design, Service Transition, Service Operation und Continual Service Improvement. Jede dieser Phasen adressiert spezifische Anforderungen und Herausforderungen, die für die Entwicklung und Implementierung von KI-Systemen besonders relevant sind.

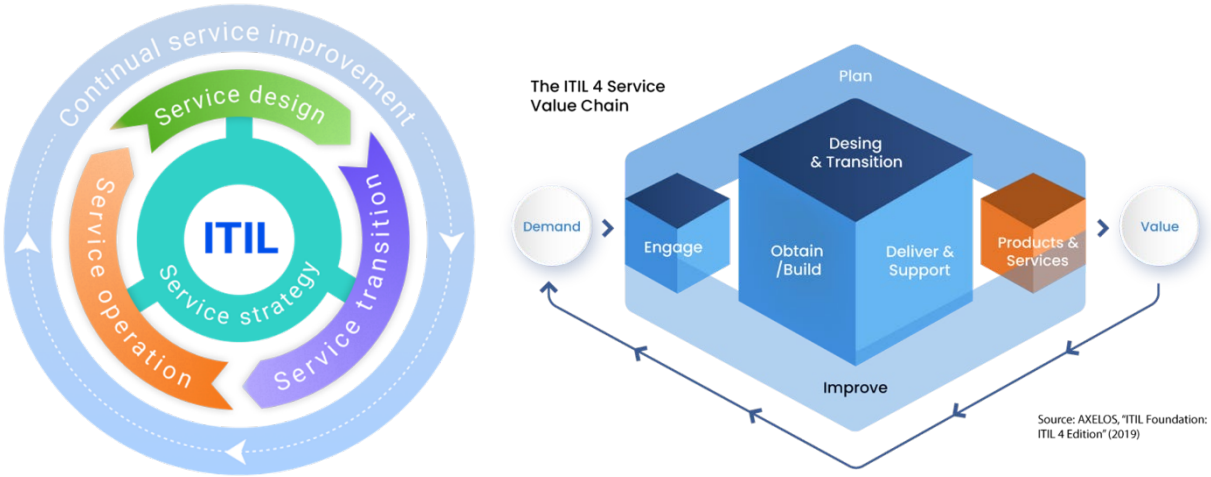


Abbildung 2: ITIL v 4 Service Value Chain

Ein erster Grund für die Nutzung von ITIL v4 liegt in der systematischen Integration verschiedener Perspektiven, welche für KI-Systeme besonders wichtig sind. KI-Produkte und Services zeichnen sich durch eine erhebliche technische Komplexität aus, kombiniert mit hohen ethischen, regulatorischen und gesellschaftlichen Anforderungen. Das ITIL-Modell ermöglicht hier eine strukturierte Bearbeitung und Integration all dieser Aspekte. So kann beispielsweise bereits in der „Service Strategy"-Phase strategisch festgelegt werden, welche regulatorischen Vorgaben (wie der AI-Act) im späteren Entwicklungs- und Betriebsprozess berücksichtigt werden müssen.

Zudem ist ITIL v4 bekannt dafür, klare und definierte Schnittstellen zwischen strategischen, taktischen und operativen Ebenen zu schaffen. Gerade bei KI-Projekten, die typischerweise verschiedene Stakeholder involvieren – von Entscheidungsträgern über Entwickler und Datenanalysten bis hin zu Endanwendern und Aufsichtsbehörden –, ist es unerlässlich, eine klare Kommunikation und Zusammenarbeit über alle Ebenen hinweg sicherzustellen. Durch den Lifecycle-Ansatz können Anforderungen und Erwartungen aller Beteiligten systematisch aufgenommen und kontinuierlich berücksichtigt werden.

Ein weiterer entscheidender Grund für die Wahl des ITIL-Modells liegt in seiner Flexibilität und Anpassungsfähigkeit. KI-Projekte, insbesondere solche in kritischen Infrastrukturen und in der Strafverfolgung, erfordern iterative Prozesse, bei denen kontinuierliche Anpassungen und Optimierungen unverzichtbar sind. Das ITIL v4 Framework unterstützt explizit eine agile Vorgehensweise, in der kontinuierliche Verbesserungen (Continual Service Improvement) integraler Bestandteil sind. Diese Eigenschaft ermöglicht es, auf dynamische Herausforderungen, wie sich ändernde regulatorische Anforderungen oder technologische Fortschritte, rasch und effizient zu reagieren.

Zusätzlich bietet ITIL v4 umfangreiche Werkzeuge und Praktiken zur Risikosteuerung und Compliance-Überwachung. Angesichts der hohen regulatorischen Anforderungen an KI-Systeme – insbesondere durch den AI-Act der Europäischen Union – bietet ITIL v4 strukturierte Methoden, um Risiken frühzeitig zu erkennen, zu dokumentieren und effektiv zu managen. Der AI-Lifecycle verbindet sich hierbei optimal mit dem Risikoregister, um sicherzustellen, dass sämtliche potenziellen Risiken kontinuierlich beobachtet und adressiert werden.

Weiterhin wurde erkannt, dass KI-Systeme nicht nur technische Herausforderungen beinhalten, sondern auch tiefgreifende ethische und gesellschaftliche Implikationen mit sich bringen. Die kontinuierliche Einbindung von ethischen Überprüfungen und Stakeholder-Feedback in allen Phasen des Lifecycle-Modells ist daher unverzichtbar. ITIL v4 bietet hierfür explizite Mechanismen wie regelmäßige Reviews, Audits und Governance-Prozesse, welche sicherstellen, dass ethische Überlegungen sowie öffentliche Akzeptanz stets berücksichtigt und optimiert werden.

Schließlich gewährleistet die Verwendung der bewährten und international anerkannten ITIL v4-Logik, dass Organisationen bei der Entwicklung und Implementierung ihrer KI-Systeme regulatorisch konform und operativ effizient agieren. Durch die Nutzung eines etablierten Standards wie ITIL v4 erhöht sich die Transparenz der Prozesse, was wiederum das Vertrauen von Stakeholdern, Nutzern und der breiten Öffentlichkeit stärkt.

Zusammenfassend wurde die Entscheidung für das ITIL v4 Service Lifecycle-Modell getroffen, weil dieses Framework systematisch, bewährt und zugleich flexibel genug ist, um den komplexen Anforderungen von KI-Projekten gerecht zu werden. Es bietet eine klare Struktur für die Zusammenarbeit verschiedener Stakeholder, die Berücksichtigung regulatorischer Anforderungen sowie die kontinuierliche Verbesserung und Anpassung von KI-Produkten und -Services. Damit stellt es eine ideale Grundlage dar, um KI-Systeme nachhaltig und verantwortungsvoll zu entwickeln, einzusetzen und langfristig erfolgreich zu betreiben.

AI Product Lifecycle: Schritte

Dieser sogenannte AI Product Lifecycle basiert in seiner Logik und Struktur auf bewährten Service-Management-Methoden wie dem ITIL v4 Framework. ITIL v4 stellt eine umfangreiche Sammlung von Best Practices dar, die Organisationen dabei helfen, Services systematisch und effizient zu entwickeln, bereitzustellen und kontinuierlich zu optimieren. Für die Anwendung auf KI-Produkte wurde dieser Lifecycle jedoch spezifisch erweitert und angepasst, um den einzigartigen Herausforderungen der KI-Technologie gerecht zu werden.

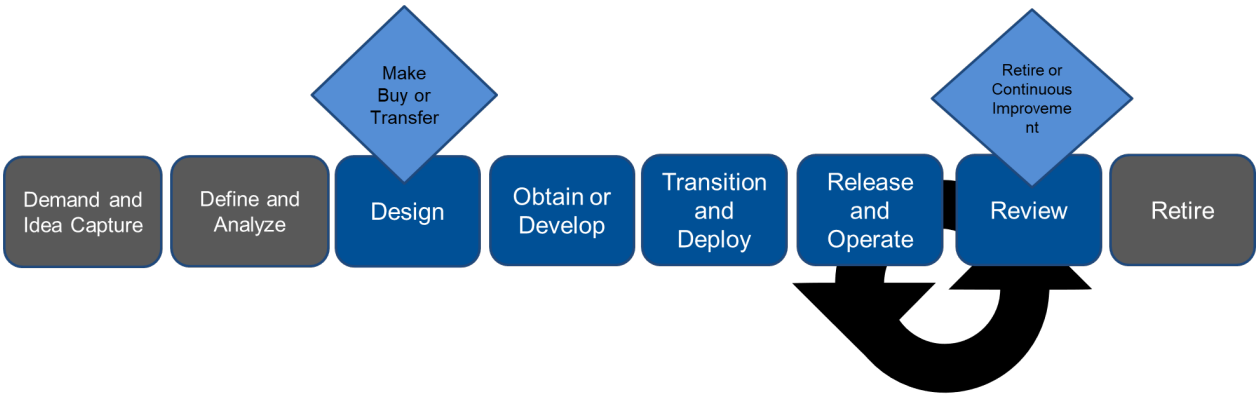


Abbildung 3: AI Product Lifecycle

Der AI Product Lifecycle ist ein iterativer und zugleich strukturierter Prozess, bestehend aus mehreren aufeinanderfolgenden und miteinander verbundenen Phasen: Demand and Idea Capture, Define and Analyze, Design, Obtain or Develop, Transition and Deploy, Release and Operate, Review und schließlich Retire. Jede dieser Phasen umfasst konkrete Aktivitäten und Aufgabenstellungen, um sicherzustellen, dass alle relevanten Anforderungen, insbesondere regulatorische Vorgaben wie der AI-Act, berücksichtigt werden. Ziel ist es, KI-Systeme nicht nur technisch robust, sondern auch ethisch vertretbar, sozial akzeptabel und rechtlich konform zu gestalten.

1. Demand and Idea Capture

- Identifizierung und Dokumentation operativer Herausforderungen oder Chancen, bei denen KI Mehrwert bietet.
- Einbindung relevanter Stakeholder (z. B. Anwender, Analysten, politische Entscheidungsträger), um Anforderungen an eine KI-Lösung zu sammeln.
- Erste Bewertung von Machbarkeit, regulatorischer Compliance und Übereinstimmung mit organisatorischen Zielen.

2. Define and Analyze

- Detaillierte Problemanalyse zur Definition der Rolle von KI bei der Problemlösung.
- Festlegung messbarer Ziele (z.B. schnellere Reaktionszeiten, weniger Fehlalarme, verbesserte Ermittlungsfähigkeit).
- Durchführung ethischer, rechtlicher und datenschutzbezogener Prüfungen, um regulatorische Anforderungen einzuhalten.
- Risikoabschätzung, einschließlich potenzieller Verzerrungen (Bias), Missbrauch oder unbeabsichtigter Folgen.

3. Design

- Entwicklung einer Lösungsarchitektur, angepasst an spezifische Anforderungen (z.B. Sicherheit, Skalierbarkeit, Integration bestehender Systeme).
- Entscheidung zur „Make-Buy-or-Transfer“-Strategie:
    - **Make:** Eigenentwicklung mit internen Ressourcen.

- **Buy:** Einkauf einer bestehenden Lösung.
- **Transfer:** Kooperation mit anderen Organisationen oder Anpassung vorhandener Lösungen.

- Erstellung eines detaillierten KI-Produktdesign-Dokuments einschließlich Anforderungen zur Compliance, Risikominderung und Betriebsvorgaben.

4. Obtain or Develop

- Beschaffung oder Entwicklung des KI-Produkts gemäß der gewählten Strategie.
- Iterative Tests zur Sicherstellung der Einhaltung funktionaler, ethischer und operativer Anforderungen.
- Training der KI-Modelle auf spezifischen Datensätzen, um Datenschutz und ethische Standards zu gewährleisten.
- Validierung des KI-Systems mit Stakeholdern durch Simulationen oder Pilotprojekte.

5. Transition and Deploy

- Vorbereitung der Infrastruktur und des Betriebs auf die Implementierung (Hardware, Software, Sicherheit).
- Durchführung umfassender Schulungen der Endnutzer (z.B. Ermittler, Analysten).
- Schrittweise Einführung des KI-Produkts durch Pilotphasen vor der vollständigen Skalierung.
- Monitoring der Erstimplementierung zur Bewertung von Leistung, Fehlern und unerwarteten Folgen.

6. Release and Operate

- Fortlaufende Wartung des KI-Produkts, einschließlich Aktualisierungen der Modelle, Leistungsoptimierungen und Sicherheitsupdates.
- Etablierung von Governance-Prozessen zur Überwachung der KI-Ergebnisse, einschließlich Genauigkeit, Verzerrungserkennung und ethischer Konformität.
- Einrichtung von Feedback-Loops zur kontinuierlichen Verbesserung basierend auf Nutzerrückmeldungen und operativen Daten.

7. Review

- Periodische Bewertungen der Effektivität des KI-Tools hinsichtlich der ursprünglichen strategischen Ziele.
- Wiederholte Risikoabschätzungen, um neuen Herausforderungen wie veränderten Datenschutzgesetzen oder sich entwickelnden Bedrohungen gerecht zu werden.
- Entscheidungspunkt zur weiteren Optimierung des bestehenden Systems, Integration neuer Funktionen oder zur Ausphasung des Tools.

8. Retire

- Entwicklung eines klaren Plans zur Außerbetriebnahme, falls das System obsolet wird (z.B. durch neue regulatorische Vorgaben oder bessere technologische Alternativen).

- Sichere Archivierung historischer Daten unter Einhaltung gesetzlicher Datenhaltungsvorschriften.
- Entfernung des Zugriffs auf das Tool und sichere Beendigung aller Abhängigkeiten.
- Dokumentation der gewonnenen Erkenntnisse, um zukünftige KI-Initiativen zu unterstützen.

Zusätzliche Überlegungen speziell für den KI-Einsatz bei Strafverfolgungsbehörden:

- **Ethische Überwachung:** Kontinuierliche ethische Begleitung zur Vermeidung von Missbrauch.
- **Stakeholder-Engagement:** Regelmäßige Einbeziehung politischer Entscheidungsträger, der Öffentlichkeit und interner Nutzer.
- **Audits und Compliance:** Periodische Audits zur Einhaltung rechtlicher und ethischer Standards.
- **Risikomanagement:** Erstellung von Notfallplänen zur Bewältigung möglicher Fehlfunktionen, Missbrauchsfälle oder negativer öffentlicher Wahrnehmung.

AI Product Lifecycle mit Beispiel

Insgesamt bietet der AI Product Lifecycle einen klar definierten Rahmen, innerhalb dessen KI-basierte Lösungen von der ersten Idee bis zur finalen Außerbetriebnahme nachhaltig gemanagt werden können. Durch die Verwendung der bewährten ITIL v4-Methodik, ergänzt um spezifische Elemente zur Berücksichtigung der besonderen Herausforderungen und Risiken von KI-Technologien, ist es gelungen, ein robustes und praxisorientiertes Lifecycle-Modell zu entwickeln. Dieses Modell erlaubt Organisationen nicht nur, technologische Innovationen effizient umzusetzen, sondern auch gesellschaftliches Vertrauen in den verantwortungsvollen Umgang mit KI-Systemen langfristig zu stärken.

Holistischer KI-Risikomanagement-Ansatz

Durch diese enge Verzahnung von ***AI Product Lifecycle***, ***DST-Framework*** und ***Risikomanagement mit dem Risikoregister*** wird gewährleistet, dass KI-Systeme nicht nur effizient und leistungsfähig, sondern auch sicher, erklärbar und regelkonform entwickelt und betrieben werden – insbesondere in sensiblen, hoch-riskanten Bereichen mit staatlicher Beteiligung oder kritischer Infrastruktur.

Das **AI Service Lifecycle** bildet den übergeordneten Rahmen für den operativen Betrieb und die kontinuierliche Verbesserung eines KI-Dienstes.

Das **DST-Framework** unterstützt den Entwicklungsprozess und hilft, den gesamten Projektverlauf von der Idee bis zum Deployment strukturiert zu planen und durchzuführen.

Das **Risikoregister** ergänzt beide, indem es spezifisch Risiken identifiziert und managt – unabhängig davon, ob es um den Entwicklungsprozess (im DST-Framework) oder den laufenden Betrieb (im AI Service Lifecycle) geht.

Während das DST-Framework vor allem bei der **Planung und Umsetzung** eines KI-Projekts hilft, sorgt der AI Service Lifecycle für einen **stabilen Betrieb**.

Das **Risikoregister** ist dabei der **gemeinsame Nenner**, der in beiden Bereichen dafür sorgt, dass Risiken frühzeitig erkannt, bewertet und durch entsprechende Maßnahmen gemindert werden.

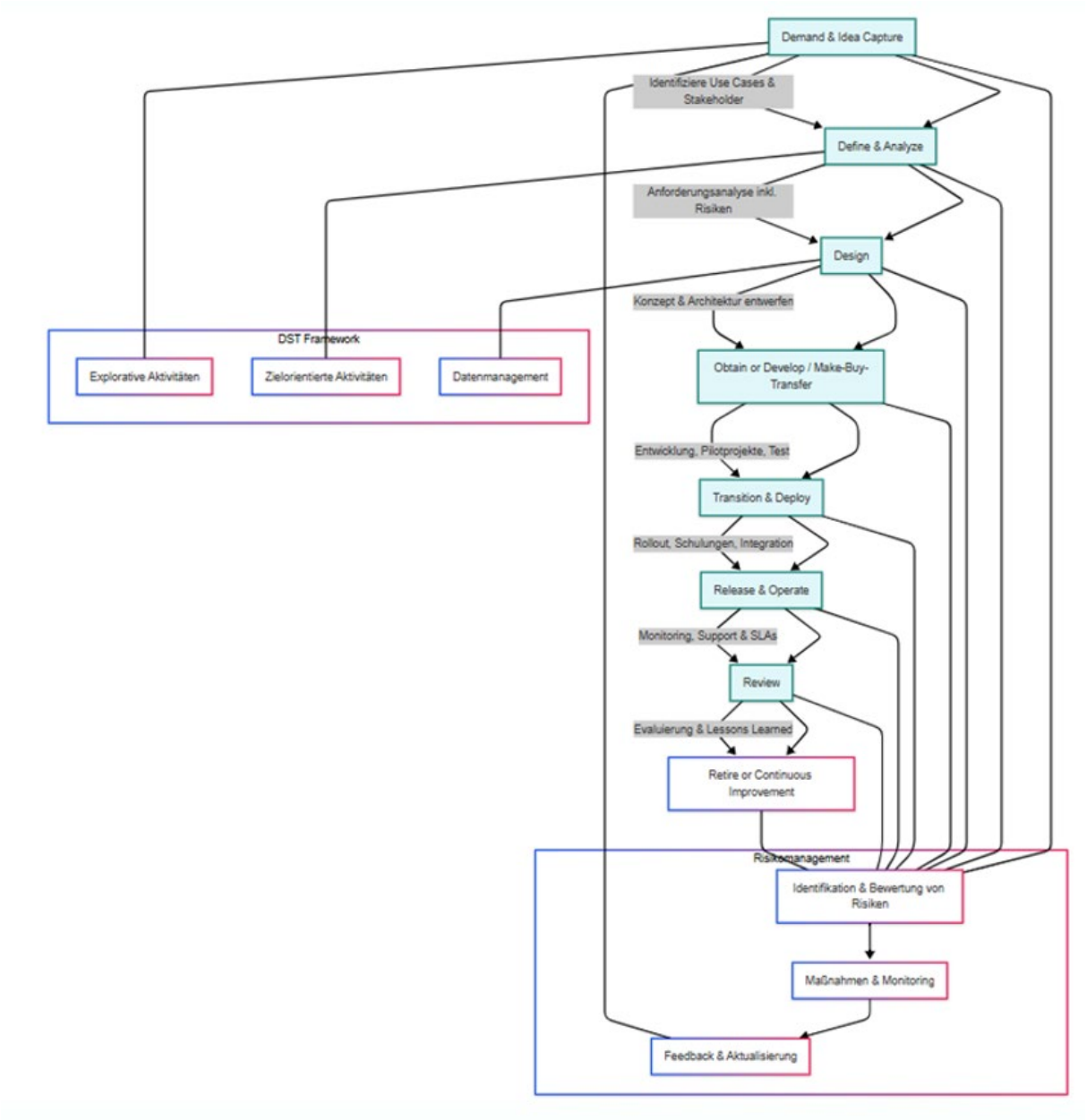


Abbildung 4: Kombination der Ergebnisse

Der gesamte Prozess gliedert sich in 5 Hauptphasen:

1. **Initiale Planung & Vorbereitung**

2. **Entwicklung und Design**
3. **Implementierung & Deployment**
4. **Betrieb & Überwachung**
5. **Review & Kontinuierliche Verbesserung**

In jeder Phase kommen unterschiedliche Elemente zum Einsatz:

- **DST-Framework:** Dient vor allem zur Planung, Modellierung und Evaluation von KI-Projekten.
- **AI Service Lifecycle:** Schafft den Rahmen für den Betrieb und die Servicebereitstellung.
- **Risikoregister:** Ist das zentrale Instrument zur Identifikation, Bewertung und Steuerung von Risiken, das in allen Phasen integriert wird.

Phase	Beschreibung	AI Service Lifecycle	DST-Framework	Risikoregister
1. Initiale Planung & Vorbereitung	Definition von Zielen, Identifikation von Use Cases und erste Risiko-Erfassung.	Strategische Planung und Service Strategy	Ziel-Exploration, Identifikation von Datenquellen	Erste Risikoidentifikation, Festlegung von Kategorien
2. Entwicklung & Design	Ausarbeitung des technischen Konzepts, Modellierung und Bewertung der Anforderungen (Make/Buy/Transfer).	Service Design (Architektur, Schnittstellen)	Konzeptentwicklung, Modellierung, Evaluierung	Detaillierte Risikoanalyse, Bewertung und Maßnahmenplanung
3. Implementierung & Deployment	Umsetzung der Lösung, Durchführung von Pilotprojekten, Integration in bestehende Systeme.	Service Transition (Rollout, Schulungen)	Umsetzung und Testing, Pilotphase	Kontinuierliche Aktualisierung und Monitoring der Risiken
4. Betrieb & Monitoring	Betrieb der KI-Lösung, kontinuierliches Monitoring der Performance, Datenqualität und Einhaltung von SLAs.	Service Operation (Betrieb, Support, Updates)	Laufendes Performance-Monitoring, Feedbackprozesse	Fortlaufendes Risiko-Monitoring, Incident Management
5. Review & Kontinuierliche Verbesserung	Regelmäßige Evaluation, Retraining, Anpassung und Optimierung der Lösung auf Basis	Continual Service Improvement	Lessons Learned, Evaluierung, Anpassung des Modells	Revision und Anpassung des Risikoregisters, Feedbackintegration

	von Lessons Learned.			

Tabelle 5: Gegenüberstellung der Prozessphasen

Data Science Trajectories Framework

Einführung

Das Data Science Trajectories (DST)-Framework bietet eine umfassende Struktur für das Verständnis und die Durchführung von Data-Science-Projekten und führt Praktiker durch das komplexe Zusammenspiel von Aktivitäten, die erforderlich sind, um verwertbare Erkenntnisse zu gewinnen und Unternehmensziele zu erreichen. Durch die Kategorisierung von Prozessen in zielgerichtete Aktivitäten, explorative Aktivitäten und Datenmanagement-Aktivitäten gewährleistet das Framework einen ganzheitlichen Ansatz zur Bewältigung datenwissenschaftlicher Herausforderungen.

Der DST-Rahmen baut auf den Grundsätzen des branchenübergreifenden Standardprozesses für Data Mining (CRISP-DM) auf und berücksichtigt gleichzeitig moderne Fortschritte in der Datenwissenschaft, wie die Erkundung neuartiger Datenquellen, die Verwendung synthetischer Daten und die Entwicklung innovativer Produkte und Erzählungen auf der Grundlage von Erkenntnissen. Jede Kategorie von Aktivitäten ist so konzipiert, dass sie verschiedene Aspekte des Data-Science-Lebenszyklus abdeckt, von der Festlegung von Geschäftszielen und der Beschaffung von Daten bis hin zur Entwicklung von Modellen, dem Einsatz von Lösungen und der Erkundung ungenutzter Möglichkeiten.

Mit seinem Schwerpunkt auf Modularität und Flexibilität ermöglicht das DST-Framework Organisationen die Anpassung an dynamische Umgebungen und gewährleistet, dass datengesteuerte Initiativen relevant, ethisch vertretbar und wirkungsvoll bleiben. Darüber hinaus unterstreicht das Rahmenwerk die Bedeutung von Governance und Risikomanagement, insbesondere im Zeitalter der künstlichen Intelligenz (KI), in dem Vorurteile, ethische Bedenken und Datenschutzfragen in jeder Phase eines Projekts sorgfältig berücksichtigt werden müssen.

Dieses Dokument bietet einen detaillierten Einblick in jede Komponente des DST-Rahmens mit ausführlichen Beschreibungen, Inputs, Outputs, Kriterien zur Fortschrittsmessung und einer Analyse der allgemeinen KI-Risiken. Anhand dieses Rahmens können Data-Science-Teams sicherstellen, dass ihre Projekte nicht nur technisch robust sind, sondern auch mit strategischen Zielen und gesellschaftlichen Werten übereinstimmen.

Kategorien

Explorative Aktivitäten

Die explorativen Aktivitäten umfassen ergebnisoffene Untersuchungen und Ideenfindung, um Möglichkeiten für datengestützte Erkenntnisse, Innovationen und Wertschöpfung zu ermitteln. Im Gegensatz zu zielgerichteten Aktivitäten lassen explorative Schritte Flexibilität, Kreativität und Anpassungsfähigkeit zu und generieren dabei oft neue Fragen, Hypothesen und potenzielle Anwendungsfälle.

1. **Erkundung der Ziele**:
   Identifizierung potenzieller Ziele, die mit Hilfe der Datenwissenschaft erreicht werden

können, oft durch Brainstorming und Trendanalysen.

2. **Erkundung von Datenquellen**:
   Es geht darum, neue Datenquellen zu entdecken und ihre Relevanz, Qualität und Zugänglichkeit zu bewerten.

3. **Erforschung des Datenwerts**:
   Bewertet den Geschäftswert oder die potenziellen Erkenntnisse, die aus Daten abgeleitet werden können, und berechnet die Machbarkeit und den voraussichtlichen Nutzen.

4. **Erforschung der Ergebnisse**:
   Untersucht die Ergebnisse und bestimmt, wie sie auf Unternehmensstrategien und -ziele angewendet werden können.

5. **Narrative Erkundung**:
   Konzentriert sich auf die Entwicklung von Datengeschichten, um Ergebnisse effektiv zu kommunizieren, oft unter Einbeziehung von Visualisierungen für mehr Klarheit und Wirkung.

6. **Produktuntersuchung**:
   Definiert potenzielle Produkte oder Dienstleistungen, die aus den Datenerkenntnissen hervorgehen können, z. B. neue Anwendungen oder Kundenlösungen.

Diese Aktivitäten sind von Natur aus iterativ und nicht linear und ermöglichen es den Teams, sich auf der Grundlage von Entdeckungen und Erkenntnissen neu auszurichten, und helfen Unternehmen, innovativ zu bleiben und auf neue Datentrends zu reagieren.

Zielgerichtete Aktivitäten

Zielorientierte Aktivitäten bestehen aus Aktivitäten, die aus dem CRISP-DM-Rahmen (Cross-Industry Standard Process for Data Mining) abgeleitet sind, der weithin als strukturierte Methodik für Data Science- und Data Mining-Projekte anerkannt ist. Zielorientierte Aktivitäten konzentrieren sich auf das Erreichen spezifischer, messbarer Geschäftsziele durch einen systematischen und oft linearen Ansatz.

1. **Geschäftsverständnis**:
   Definiert die Ziele des Projekts im geschäftlichen Kontext und stellt die Übereinstimmung mit den Unternehmenszielen sicher.

2. **Verstehen der Daten**:
   Bewertung der Relevanz, der Qualität und der Grenzen der Daten zur Erfüllung der Projektanforderungen.

3. **Datenaufbereitung**:
   Bereinigung und Umwandlung von Daten in ein für die Analyse geeignetes Format, Verbesserung der Datenqualität und -konsistenz.

4. **Modellierung**:
   Wendet maschinelles Lernen oder statistische Methoden an, um Erkenntnisse oder Vorhersagen zu generieren, indem er Modelle auswählt und abstimmt, um sie an die Daten anzupassen.

5. **Evaluierung**:
   Misst die Leistung des Modells und die Übereinstimmung mit den Unternehmenszielen und stellt sicher, dass die Ergebnisse gültig, zuverlässig und umsetzbar sind.

6. **Einsatz**:
   Integration des endgültigen Modells in das Unternehmensumfeld, damit die Beteiligten die Ergebnisse effektiv nutzen können.

Diese Aktivitäten sind sequenziell organisiert, können jedoch iterative Schleifen aufweisen, insbesondere wenn eine Neubewertung oder weitere Verfeinerung erforderlich ist. Jede Phase liefert Ergebnisse, die die nachfolgenden Schritte leiten und dazu beitragen, die gewünschten geschäftlichen Auswirkungen zu erzielen.

Datenmanagementaktivitäten

Die Datenverwaltung konzentriert sich auf die systematische Handhabung, Organisation und Zugänglichkeit von Daten, die für eine effektive Analyse und die langfristige Nachhaltigkeit von datengesteuerten Lösungen entscheidend sind. Diese Aktivitäten stellen sicher, dass die Daten für die Analyse verfügbar, strukturiert und zuverlässig sind, und berücksichtigen gleichzeitig Aspekte wie Datenschutz, Sicherheit und Infrastruktur.

1. **Datenerfassung**:
   Das Sammeln relevanter Daten aus verschiedenen Quellen, entweder durch Erfassung, Extraktion oder Erwerb.

2. **Daten-Simulation**:
   Erzeugt synthetische Daten durch Simulationen, besonders nützlich für die Modellierung von Szenarien, in denen reale Daten knapp oder nicht verfügbar sind.

3. **Datenarchitektur**:
   Dazu gehört der Entwurf einer logischen und physischen Struktur für die Datenspeicherung und den Datenzugriff, die eine effiziente Verarbeitung und Integration unterstützt.

4. **Datenfreigabe**:
   Bereitstellung von Daten zur Analyse oder gemeinsamen Nutzung, in der Regel über Datenbanken, APIs oder Schnittstellen, die die Zusammenarbeit und die Benutzerfreundlichkeit erleichtern.

Diese Aktivitäten sind für jedes Data-Science-Projekt von grundlegender Bedeutung, da sie sicherstellen, dass der Lebenszyklus der Daten von der Erfassung bis zur Verteilung unter Berücksichtigung von Qualität, Compliance und Skalierbarkeit aufrechterhalten wird. Datenmanagement-Aktivitäten sind für den Aufbau einer soliden, anpassungsfähigen Dateninfrastruktur, die laufende Analysen und die Skalierbarkeit zukünftiger Projekte unterstützt, unerlässlich.

Komponenten

Explorative Komponenten

Zielsetzung Exploration

Kategorie	Sondierungstätigkeiten
ID	EX-01
Beschreibung	Identifizierung und Definition übergeordneter Ziele, die mit Hilfe von Data Science erreicht werden können. Dazu gehören das Brainstorming potenzieller Anwendungsfälle, die Analyse von Trends und die Abstimmung der Ziele mit umfassenderen Unternehmenszielen.
Eingaben	• Geschäftliche Ziele • Markttrends • Wettbewerbslandschaft • Benutzer-Feedback
Ausgänge	Übergeordnete Ziele der Datenwissenschaft Prioritätenliste der potenziellen Projekte Dokument zur Zielausrichtung
Messung des Fortschritts	• Fertigstellung der Zieldefinition • Überprüfung und Genehmigung durch die Interessengruppen • Durchführbarkeitsanalyse für die vorgeschlagenen Ziele
Allgemeine AI-Risiken	Unrealistische Erwartungen an KI-Fähigkeiten Risiko der Verfolgung von Zielen, die technisch nicht machbar sind Ethische Fragen, wenn Ziele mit gesellschaftlichen Werten kollidieren

Tabelle 6: Zielsetzung Exploration

Organisatorische Ziele identifizieren

Verstehen der übergeordneten strategischen Ziele des Unternehmens, um die Ausrichtung der Data-Science-Bemühungen an den allgemeinen Geschäftsprioritäten sicherzustellen.

- **Inputs**:
  - Strategische Pläne, organisatorische Visionen und Leitbilder
  - Aktuelle Geschäftsleistungskennzahlen und KPIs
  - Befragung von Interessenvertretern und Feedback
- **Aktivitäten**:
  - Durchführung von Interviews mit Interessenvertretern zur Erfassung strategischer Ziele
  - Überprüfung der organisatorischen Vision, des Auftrags und der KPIs
  - Identifizierung von Herausforderungen und Möglichkeiten, bei denen die Datenwissenschaft einen Mehrwert bieten kann
- **Outputs**:
  - Dokument zur strategischen Ausrichtung
  - Priorisierte Liste von Organisationszielen

Brainstorming für potenzielle Anwendungsfälle

Erstellen Sie eine Liste potenzieller datengesteuerter Projekte, die sich mit identifizierten organisatorischen Herausforderungen oder Chancen befassen.

- **Inputs:**
  - Identifizierte organisatorische Herausforderungen und Ziele
  - Branchenberichte, Fallstudien und Trends in der Datenwissenschaft
  - Funktionsübergreifender Team-Input
- **Aktivitäten**:
  - Organisation von Brainstorming-Sitzungen mit funktionsübergreifenden Teams
  - Erforschung branchenspezifischer Anwendungen der Datenwissenschaft
  - Einsatz von Design Thinking oder anderen Methoden, um innovative Anwendungsfälle zu finden

- **Outputs:**
    - Anwendungsfall-Ideenliste
    - Geordnete Anwendungsfälle

Bewertung der Durchführbarkeit von Anwendungsfällen

Bewerten Sie die technische und betriebliche Machbarkeit jedes Anwendungsfalls, um unpraktische Ideen herauszufiltern.

- **Inputs:**
    - Datenverfügbarkeit und Qualitätsberichte
    - Technische Ressourcen und Infrastruktur
    - Geschätztes Budget und Zeitplan für die Umsetzung
- **Aktivitäten**:
    - Bewertung der Verfügbarkeit und Qualität der für jeden Anwendungsfall erforderlichen Daten
    - Bewerten Sie die Komplexität der technischen Umsetzung
    - Analyse des Ressourcenbedarfs, z. B. in Bezug auf Zeit, Fachwissen und Infrastruktur
- **Outputs:**
    - Berichte über Durchführbarkeitsanalysen
    - Auswahlliste der realisierbaren Anwendungsfälle

Abgleich von Anwendungsfällen mit Unternehmenszielen

Ordnen Sie die in die engere Wahl gezogenen Anwendungsfälle spezifischen Geschäftszielen zu, um messbare Auswirkungen zu gewährleisten.

- **Inputs:**
    - In die engere Wahl gezogene Anwendungsfälle und Machbarkeitsberichte
    - Unternehmensziele und strategische Prioritäten
    - Stakeholder-Feedback zu den Auswirkungen auf das Geschäft

- **Aktivitäten**:
    - Erstellen einer Mapping-Matrix zwischen Anwendungsfällen und Geschäftszielen
    - Definieren Sie messbare Erfolgskriterien für jeden Anwendungsfall
    - Sicherstellung des Feedbacks und der Zustimmung der Interessengruppen für ausgewählte Anwendungsfälle

- **Outputs:**
    - Abgleichsmatrix, die Ziele auf Anwendungsfälle abbildet
    - Genehmigte Liste der Ziele der Datenwissenschaft

Priorisierung der Ziele

Ordnen Sie die identifizierten Ziele nach ihrer potenziellen Wirkung, Durchführbarkeit und Übereinstimmung mit den Prioritäten der Organisation.

- **Inputs:**
    - Liste der genehmigten Ziele und Anwendungsfälle
    - Bewertungskriterien für Auswirkungen, Durchführbarkeit und Anpassung
    - Feedback von Interessengruppen und Teambesprechungen

- **Aktivitäten**:
    - Entwicklung eines Bewertungsrahmens für Auswirkungen, Durchführbarkeit und Anpassung
    - Durchführung von Workshops zur Prioritätensetzung mit Interessengruppen
    - Fertigstellung einer nach Prioritäten geordneten Roadmap für Data-Science-Initiativen

- **Outputs:**
    - Priorisierte Liste von Zielen
    - Fahrplan für Datenwissenschaft

Ziele dokumentieren und kommunizieren

Dokumentieren Sie die Ziele klar und kommunizieren Sie sie an alle relevanten Interessengruppen, um eine Abstimmung und ein gemeinsames Verständnis sicherzustellen.

- **Inputs:**
  - Priorisierte Ziele und Fahrplan
  - Kommunikationsmittel und Vorlagen (z. B. Präsentationen, Berichte)
  - Stakeholder-Feedback und Notizen zur Angleichungssitzung

- **Aktivitäten**:
  - Erstellen Sie eine Vorlage für eine formale Zieldokumentation
  - Entwicklung von Präsentationen oder visuellen Hilfsmitteln zur Verbreitung des Fahrplans
  - Veranstalten Sie ein Treffen zur Abstimmung der Ziele mit den Beteiligten, um die Ziele festzulegen und zu kommunizieren.

- **Outputs:**
  - Bericht zur Zieldokumentation
  - Materialien für die Kommunikation mit Interessengruppen
  - Ausrichtungsgenehmigungsprotokoll

Erkundung von Datenquellen

Kategorie	Sondierungstätigkeiten
ID	EX-02
Beschreibung	Es geht darum, potenzielle Datenquellen zu entdecken und zu bewerten, die wertvolle Erkenntnisse liefern können. Dazu gehört die Bewertung von Relevanz, Qualität, Zugänglichkeit und Kosten der Daten.
Eingaben	• Liste der potenziellen Datenquellen • Werkzeuge zur Datenerfassung • Kriterien für die Bewertung der Datenqualität
Ausgänge	Identifizierte Datenquellen Plan zur Datenerfassung Bericht über die Datenrelevanz
Messung des Fortschritts	• Identifizierung der relevanten Quellen • Erste Datenerfassung • Qualitätsbewertung und Validierung von Quellen
Allgemeine AI-Risiken	Verwendung von unzuverlässigen oder verzerrten Datenquellen Begrenzte Datenverfügbarkeit beeinträchtigt die Zuverlässigkeit des Modells Potenzielle Risiken für die Privatsphäre im Zusammenhang mit bestimmten Datentypen

Tabelle 7: Erkundung von Datenquellen

Identifizierung potenzieller Datenquellen

Ermitteln Sie mögliche Datenquellen, die zur Unterstützung der festgelegten Geschäftsziele oder Anwendungsfälle genutzt werden können. Dabei werden sowohl interne als auch externe Quellen untersucht, darunter Datenbanken, APIs, öffentliche Datensätze und Drittanbieter.

- **Inputs:**
  - Unternehmensziele und Zielsetzungen
  - Branchen-Benchmarks und Fallstudien
  - Aktuelle Dateninfrastruktur und bestehende Datensätze
- **Aktivitäten**:
  - Durchsicht der Datenbestände von Unternehmen auf relevante Datensätze
  - Recherche öffentlich zugänglicher Datensätze und APIs
  - Bewertung von Drittanbietern von Daten auf Zuverlässigkeit und Relevanz
- **Outputs:**
  - Erste Liste der potenziellen Datenquellen
  - Inventar des Datenspeichers
  - Bericht über die Ermittlung von Datenquellen

Bewertung der Relevanz und Qualität von Daten

Bewerten Sie die Relevanz, Qualität und Vollständigkeit der ermittelten Datenquellen, um sicherzustellen, dass sie den Projektanforderungen entsprechen. Dazu gehört die Überprüfung der Daten auf Gültigkeit, Genauigkeit, Aktualität und mögliche Verzerrungen.

- **Inputs:**
  - Erste Liste der potenziellen Datenquellen
  - Kriterien für die Bewertung der Datenqualität (z. B. Genauigkeit, Vollständigkeit, Konsistenz)
  - Anforderungen an den geschäftlichen Anwendungsfall

- **Aktivitäten**:
    - Durchführen einer vorläufigen Datenprofilierung für jede Quelle
    - Bewerten Sie die Fähigkeit der Daten, die Anforderungen des Anwendungsfalls zu erfüllen
    - Identifizierung potenzieller Probleme oder Lücken in der Datenqualität

- **Outputs:**
    - Bericht zur Datenrelevanz
    - Bericht zur Bewertung der Datenqualität
    - Auswahlliste der verwendbaren Datenquellen

Bewertung der Zugänglichkeit von Daten und Berechtigungen

Bestimmen Sie, wie zugänglich die in die engere Wahl gezogenen Datenquellen sind und welche Berechtigungen oder Protokolle für den Zugriff auf sie erforderlich sind. Dies gewährleistet die Einhaltung rechtlicher, ethischer und organisatorischer Richtlinien.

- **Inputs:**
    - Auswahlliste der verwendbaren Datenquellen
    - Datenzugangsrichtlinien und -protokolle
    - Rechtliche und Compliance-Richtlinien (z. B. GDPR)

- **Aktivitäten**:
    - Überprüfen Sie die Zugriffsberechtigungen für jede Datenquelle
    - Identifizieren Sie etwaige Einschränkungen oder Compliance-Anforderungen
    - Erforderlichenfalls Vereinbarungen über die gemeinsame Nutzung von Daten treffen

- **Outputs:**
    - Bericht über die Zugänglichkeit von Daten
    - Checkliste zur Einhaltung der Vorschriften
    - Liste der zugänglichen Datenquellen

Zuordnung von Datenquellen zu Anwendungsfällen

Abstimmung der ermittelten und zugänglichen Datenquellen auf bestimmte Anwendungsfälle oder Ziele. Dadurch wird sichergestellt, dass die ausgewählten Datenquellen die notwendigen Inputs für die Erreichung der Geschäftsziele liefern.

- **Inputs:**
    - Auswahlliste der zugänglichen Datenquellen
    - Anforderungen an den Anwendungsfall
    - Datenquellenprofile (z. B. Schemata, Attribute, Formate)

- **Aktivitäten**:
    - Zuordnung von Datenattributen zu spezifischen Anwendungsfallanforderungen
    - Ermittlung potenzieller Lücken, in denen zusätzliche Daten erforderlich sein könnten
    - Validierung der Angemessenheit von Datenquellen für bestimmte Anwendungsfälle

- **Outputs:**
    - Dokument zur Zuordnung von Datenquellen zu Anwendungsfällen
    - Bericht zur Lückenanalyse
    - Liste der Datenquellen validieren

Sichern und Integrieren von Datenquellen

Sichern Sie die Daten aus ausgewählten Quellen und integrieren Sie sie in die Dateninfrastruktur des Unternehmens. Dazu gehören die Einrichtung von Datenpipelines, Speichermechanismen und die Gewährleistung des laufenden Zugriffs.

- **Inputs:**
    - Liste der validierten Datenquellen
    - Datenintegrationswerkzeuge und -plattformen (z. B. ETL-Pipelines, APIs)
    - Datenmanagement und Sicherheitsrichtlinien

- **Aktivitäten**:
    - Einrichtung von Datenpipelines für die kontinuierliche oder Batch-Datenerfassung
    - Speicherung von Daten in geeigneten Repositories oder Systemen
    - Anwendung von Sicherheitsmaßnahmen zum Schutz sensibler Daten

- **Outputs:**
    - Integrierte Datensätze
    - Operative Datenpipelines
    - Bericht über die Einhaltung von Sicherheit und Governance

Überwachung und Pflege von Datenquellen

Kontinuierliche Überwachung der Datenquellen, um ihre Zuverlässigkeit, Relevanz und Qualität im Laufe der Zeit sicherzustellen. Erstellung von Protokollen für die Pflege von Datenquellen und die Behebung auftretender Probleme.

- **Inputs:**
    - Integrierte Datensätze und operative Pipelines
    - Datenüberwachungsinstrumente (z. B. Dashboards, Skripte zur Qualitätskontrolle)
    - Datenquellenverträge und SLAs (falls zutreffend)

- **Aktivitäten**:
    - Automatische Überwachung von Datenqualität und -verfügbarkeit einrichten
    - Durchführung regelmäßiger Audits, um die Relevanz der Daten und die Einhaltung der Vorschriften zu gewährleisten
    - Behebung von Problemen wie Datenabweichungen, veraltete Quellen oder Zugriffsunterbrechungen

- **Outputs:**
    - Dashboard zur Datenüberwachung
    - Regelmäßige Berichte über die Datenqualität
    - Protokolle zur Problemlösung

Datenwert-Exploration

Kategorie	Sondierungstätigkeiten
ID	EX-03
Beschreibung	Konzentriert sich auf die Einschätzung des potenziellen Werts oder der Erkenntnisse, die aus Daten gewonnen werden können. Dazu gehört die Definition möglicher Ergebnisse, Wertvorstellungen und Anwendungen der Datenanalyse.
Eingaben	• Erfasste Daten • Analyse des Geschäftsfalls • Analyse historischer Daten • Mögliche Erkenntnisse
Ausgänge	Bewertung des Unternehmenswertes Prognostizierte Erkenntnisse und Anwendungsfälle Machbarkeitsbericht
Messung des Fortschritts	• Identifizierung der erreichbaren Erkenntnisse • Schätzung des ROI oder anderer Benchmarks • Validierung des vorgeschlagenen Wertes durch die Interessengruppen
Allgemeine AI-Risiken	Überschätzung des Wertes aufgrund optimistischer Annahmen Übersehen potenzieller Verzerrungen, die die Erkenntnisse beeinträchtigen können Potenzieller Datenmissbrauch, wenn Erkenntnisse nicht ethisch korrekt verwaltet werden

Tabelle 8: Datenwert-Exploration

Potenzielle Wertvorteile definieren

Identifizieren Sie die Arten von Werten, die potenziell aus den Daten abgeleitet werden können, z. B. betriebliche Effizienz, Kundeneinblicke oder Umsatzmöglichkeiten. In diesem Schritt wird die Datenexploration auf die Unternehmensziele und strategischen Prioritäten abgestimmt.

- **Inputs:**
  - Geschäftsziele und Zielsetzungen
  - Anforderungen an den Anwendungsfall
  - Organisatorische Leistungsdaten und Branchenbenchmarks

- **Aktivitäten**:
  - Durchführung von Brainstorming-Sitzungen mit Geschäftsinteressenten
  - Untersuchung von Werttreibern in ähnlichen Branchen oder Bereichen
  - Identifizierung potenzieller Anwendungen (z. B. prädiktive Analysen, Optimierung)

- **Outputs:**
  - Erste Liste von Wertvorschlägen
  - Zuordnung von Geschäftszielen zu Datenmöglichkeiten

Analyse historischer Daten zur Gewinnung von Erkenntnissen

Verwenden Sie historische oder vorhandene Daten, um Muster, Trends oder Anomalien zu erkennen, die Aufschluss über den potenziellen Wert geben können. Diese Analyse bildet die Grundlage für die Entscheidung, ob die künftige Datenexploration wahrscheinlich aussagekräftige Ergebnisse liefern wird.

- **Inputs:**
  - Historische Datensätze
  - Analysetools (z. B. Visualisierungssoftware, statistische Bibliotheken)
  - Fachwissen von Wirtschaftsexperten

- **Aktivitäten**:
    - Durchführung explorativer Datenanalysen (EDA) zur Aufdeckung von Trends und Anomalien
    - Identifizierung von Beziehungen zwischen Datenattributen und Geschäftsergebnissen
    - Zusammenfassung der wichtigsten Erkenntnisse, die mit den strategischen Zielen übereinstimmen

- **Outputs:**
    - EDA-Berichte
    - Zusammenfassung der Dateneinsicht
    - Erste Hypothesen für datengesteuerte Möglichkeiten

Schätzung des quantifizierbaren Nutzens

Berechnen Sie den potenziellen Nutzen aus der Nutzung der Daten, z. B. Umsatzsteigerungen, Kosteneinsparungen oder verbesserte Kundenzufriedenheit. Dieser Schritt konzentriert sich auf die Erstellung messbarer Erfolgskriterien.

- **Inputs:**
    - Unternehmensleistungskennzahlen und KPIs
    - Identifizierte Wertvorstellungen
    - Vorläufige Erkenntnisse aus der Analyse historischer Daten

- **Aktivitäten**:
    - Entwicklung von ROI-Modellen zur Abschätzung der finanziellen Auswirkungen
    - Quantifizierung potenzieller Verbesserungen bei Schlüsselkennzahlen (z. B. Verringerung der Abwanderung)
    - Definition von Erfolgsmetriken für jedes identifizierte Wertversprechen

- **Outputs:**
    - ROI-Schätzungen
    - Quantifizierte Nutzenmodelle
    - Dokument Erfolgskriterien

Validierung der Durchführbarkeit der Wertrealisierung

Beurteilen Sie, ob der ermittelte Wert angesichts der Daten, Ressourcen und organisatorischen Beschränkungen realistischerweise erreicht werden kann. Dazu gehört auch die Identifizierung technischer, betrieblicher oder rechtlicher Hürden.

- **Inputs:**
  - ROI-Schätzungen und Erfolgskriterien
  - Datenverfügbarkeit und Qualitätsberichte
  - Bewertung der organisatorischen Ressourcen und Fähigkeiten

- **Aktivitäten**:
  - Analyse von Abhängigkeiten zur Erzielung von Mehrwert (z. B. zusätzliche Datenquellen)
  - Identifizierung von Risiken oder Beschränkungen, die die Wertrealisierung einschränken könnten
  - Validierung von Wertvorschlägen mit Interessengruppen und Fachleuten

- **Outputs:**
  - Durchführbarkeitsbericht
  - Validierte Liste von realisierbaren Wertvorschlägen
  - Dokument zur Risikobewertung

Erkundung von Szenarien für die Wertschaffung

Erstellen Sie Szenarien oder Simulationen von Anwendungsfällen, um zu untersuchen, wie der Wert der Daten unter verschiedenen Bedingungen genutzt werden kann. Dieser Schritt unterstützt die Entscheidungsfindung, indem er die praktischen Auswirkungen der Datenverwendung aufzeigt.

- **Inputs:**
  - Überprüfte Wertvorstellungen
  - Tools zur Datensimulation und historische Einblicke
  - Stakeholder-Feedback zu operativen Szenarien

- **Aktivitäten**:
    - Entwicklung von Simulationen/Modellen zur Prüfung der Wertextraktion unter verschiedenen Bedingungen
    - Bewertung der Auswirkungen hypothetischer Veränderungen (z. B. Markttrends)
    - Identifizierung optimaler Ansätze für die Wertschöpfung

- **Outputs:**
    - Szenario-Modelle
    - Pläne zur Wertsteigerung
    - Ergebnisse der Simulation/Modell

Priorisierung von Wertschöpfungsmöglichkeiten

Einstufung von Wertschöpfungsmöglichkeiten auf der Grundlage von Auswirkung, Durchführbarkeit und Übereinstimmung mit den Unternehmenszielen. So wird sichergestellt, dass die Ressourcen den vielversprechendsten Initiativen zugewiesen werden.

- **Inputs:**
    - Validierte Wertvorstellungen und Machbarkeitsbewertungen
    - ROI-Modelle und Ergebnisszenarien
    - Prioritäten der Interessengruppen und strategische Ziele

- **Aktivitäten**:
    - Entwicklung eines Bewertungsrahmens für die Einstufung von Chancen
    - Durchführung von Workshops zur Prioritätensetzung mit Interessengruppen
    - Fertigstellung einer nach Prioritäten geordneten Liste von Wertschöpfungsmöglichkeiten

- **Outputs:**
    - Geordnete Wertmöglichkeiten
    - Plan für die Ressourcenzuweisung
    - Rahmen für die Wertpriorisierung

Dokumentation und Kommunikation von Wertschöpfungsmöglichkeiten

Ermittelte Wertschöpfungsmöglichkeiten klar zu dokumentieren und sie den wichtigsten Interessengruppen zur Genehmigung und Abstimmung mitzuteilen.

- **Inputs:**
  - Priorisierte Wertschöpfungsmöglichkeiten und Ressourcenpläne
  - Kommunikationsmittel und Vorlagen (z. B. Präsentationen, Dashboards)
  - Stakeholder-Feedback zu Wertschöpfungsmöglichkeiten

- **Aktivitäten**:
  - Erstellen Sie ein formelles Dokument, das die Wertmöglichkeiten zusammenfasst.
  - Entwicklung von Visualisierungen zur Vermittlung wichtiger Erkenntnisse
  - Veranstalten Sie eine Sitzung zur Wertanpassung mit Interessengruppen

- **Outputs:**
  - Dokumentation der Wertschöpfungsmöglichkeiten
  - Materialien für die Kommunikation mit Interessengruppen
  - Ausrichtungsgenehmigungsprotokoll

Kategorie	Sondierungstätigkeiten
ID	EX-04
Beschreibung	Bewertet, wie die Ergebnisse mit den Unternehmenszielen übereinstimmen und effektiv angewendet werden können. Dazu gehört das Testen und Erforschen von Ergebnissen, um zu sehen, wie sie strategische Maßnahmen vorantreiben können.
Eingaben	• Ergebnisse der Datenanalyse • Geschäftliche Ziele • Überlegungen zum Einsatz
Ausgänge	Bericht zur Geschäftsausrichtung Empfehlungen für Maßnahmen Ergebnisse des Pilotprojekts
Messung des Fortschritts	• Grad der Übereinstimmung mit den Unternehmenszielen • Feedback von Interessengruppen • Feedback zur Pilotdurchführung
Allgemeine AI-Risiken	Diskrepanz zwischen Ergebnissen und Unternehmenszielen Mögliche unbeabsichtigte Folgen der Anwendung der Ergebnisse Ethische Bedenken hinsichtlich der Anwendung bestimmter Ergebnisse

Tabelle 9: Ergebnis-Exploration

Interpretation der ersten Ergebnisse

Analysieren Sie die aus der Datenanalyse oder Modellierung gewonnenen Ergebnisse, um ihre Bedeutung und Relevanz für die Unternehmensziele zu verstehen. In diesem Schritt geht es darum, die Rohdaten sinnvoll auszuwerten und umsetzbare Erkenntnisse zu ermitteln.

- **Inputs:**
  - Datenanalyse oder Modellierungsergebnisse
  - Geschäftsziele und Anforderungen an Anwendungsfälle
  - Statistische Werkzeuge und Visualisierungssoftware

- **Aktivitäten**:
  - eine detaillierte Überprüfung der Analyse- oder Modellierungsergebnisse durchführen
  - Erzeugen von Visualisierungen zur Interpretation von Mustern und Beziehungen
  - Zusammenfassung der wichtigsten Ergebnisse in Bezug auf die Unternehmensziele

- **Outputs:**
  - Bericht über erste Erkenntnisse
  - Visualisierung von Dashboards
  - Zusammenfassung der Ergebnisinterpretation

Validierung der Ergebnisse anhand der Ziele

Sicherstellen, dass die Ergebnisse mit den vordefinierten Geschäftszielen und Erfolgskriterien übereinstimmen. Dies beinhaltet die Prüfung auf Konsistenz, Relevanz und mögliche Verzerrungen der Ergebnisse.

- **Inputs:**
  - Erste Erkenntnisse und Ergebniszusammenfassungen
  - Erfolgskriterien und Unternehmensziele
  - Feedback von Fachleuten und/oder Interessengruppen

- **Aktivitäten**:
    - Ergebnisse mit Erfolgskriterien vergleichen
    - Identifizierung von Diskrepanzen, Verzerrungen oder Anomalien
    - Einholen von Feedback von Interessengruppen zur Bestätigung der Relevanz

- **Outputs:**
    - Validierungsbericht
    - Zusammenfassung des Stakeholder-Feedbacks
    - Liste der validierten Ergebnisse

Erkundung alternativer Erklärungen

Untersuchung alternativer Hypothesen oder Erklärungen für die Ergebnisse, um die Robustheit zu gewährleisten und das Risiko von Fehlinterpretationen zu verringern. Dieser Schritt vertieft die Analyse, indem andere mögliche Kausalfaktoren untersucht werden.

- **Inputs:**
    - Validierte Ergebnisse
    - Fachwissen und historische Daten
    - Statistische Prüfwerkzeuge

- **Aktivitäten**:
    - Sensitivitätsanalyse durchführen, um alternative Erklärungen zu testen
    - Vergleich der Ergebnisse mit historischen Trends oder Benchmarks
    - Dokumentieren Sie mögliche ursächliche Faktoren und bewerten Sie deren Wahrscheinlichkeit

- **Outputs:**
    - Bericht zur Sensitivitätsanalyse
    - Alternatives Erklärungsdokument
    - Bewertung der Robustheit

Identifizierung geschäftlicher Auswirkungen

Verknüpfen Sie die Ergebnisse mit spezifischen Geschäftsaktionen oder -entscheidungen und identifizieren Sie Chancen, Risiken oder Änderungen, die erforderlich sind, um die Erkenntnisse effektiv zu nutzen.

- **Inputs:**
    - Validierte und robuste Ergebnisse
    - Strategische Ziele und geschäftlicher Kontext
    - Feedback von Entscheidungsträgern

- **Aktivitäten**:
    - Erkenntnisse in umsetzbare geschäftliche Implikationen umwandeln
    - Ermittlung von Bereichen, in denen Entscheidungen oder Strategien verbessert werden können
    - Aufzeigen von Risiken oder Ungewissheiten im Zusammenhang mit bestimmten Maßnahmen

- **Outputs:**
    - Bericht über die Auswirkungen auf die Wirtschaft
    - Empfehlungen des Aktionsplans
    - Dokument zur Risikobewertung

Prüfung der Skalierbarkeit und Anwendbarkeit

Bewerten Sie, ob die Ergebnisse verallgemeinert oder auf andere Unternehmensbereiche übertragen werden können, um eine breitere Anwendbarkeit zu gewährleisten. In diesem Schritt wird untersucht, inwieweit die Ergebnisse in unterschiedlichen Kontexten relevant sind.

- **Inputs:**
    - Umsetzbare Erkenntnisse und Auswirkungen
    - Zusätzliche Datensätze oder Ergebnisse von Pilotprojekten
    - Operative Rückmeldungen aus Umsetzungsversuchen

- **Aktivitäten**:
    - Testergebnisse in kleinem Maßstab oder Pilotumgebungen
    - Bewertung der Konsistenz von Erkenntnissen über verschiedene Datensätze oder Bereiche hinweg
    - Einholen von Feedback zu den praktischen Herausforderungen bei der Skalierung der Erkenntnisse

- **Outputs:**
    - Bericht zur Skalierbarkeit
    - Ergebnisse des Pilotprojekts
    - Validierung der Verallgemeinerung

Kommunikation der Ergebnisse und Auswirkungen

Effektive Kommunikation der Ergebnisse, Auswirkungen und empfohlenen Maßnahmen an die Beteiligten, um eine Abstimmung und fundierte Entscheidungsfindung zu gewährleisten.

- **Inputs:**
    - Validierte Ergebnisse und geschäftliche Auswirkungen
    - Kommunikationsmittel (z. B. Berichte, Präsentationen, Dashboards)
    - Anforderungen an die Kommunikation mit den Stakeholdern

- **Aktivitäten**:
    - Entwicklung umfassender Berichte oder Dashboards, die die Ergebnisse zusammenfassen
    - Erstellen Sie Visualisierungen, die wichtige Erkenntnisse und Auswirkungen hervorheben
    - Präsentation der Ergebnisse vor den Interessenvertretern unter Berücksichtigung ihres Feedbacks

- **Outputs:**
    - Materialien für die Präsentation von Interessengruppen
    - Ergebnismitteilungsbericht
    - Zusammenfassung des Feedbacks

Überwachung und Überprüfung der Ergebnisse

Einrichtung eines Mechanismus zur kontinuierlichen Überwachung der aus den Ergebnissen abgeleiteten Maßnahmen und deren Überprüfung, wenn neue Daten oder Umstände auftreten.

- **Inputs:**
    - Ergebnisse Umsetzung der Ergebnisse
    - Überwachungsinstrumente (z. B. Dashboards, KPI-Tracker)
    - Feedback von operativen Teams

- **Aktivitäten**:
    - Aufstellung von KPIs und Überwachungsrahmen für die Verfolgung der Ergebnisse
    - Regelmäßige Überprüfung der Ergebnisse, um Abweichungen oder Verbesserungen zu ermitteln
    - Aktualisierung der Ergebnisse und Auswirkungen auf der Grundlage neuer Daten oder Erkenntnisse

- **Outputs:**
    - Überwachung des Dashboards
    - Regelmäßige Überprüfungsberichte
    - Aktualisierte Ergebnisinterpretationen

Kategorie	Sondierungstätigkeiten
ID	EX-05
Beschreibung	Es geht darum, aus den Daten überzeugende Geschichten zu entwickeln, um die Ergebnisse effektiv zu kommunizieren. Diese Phase umfasst Visualisierungen, Data Storytelling und die Erstellung von Erzählungen, die Erkenntnisse zugänglich machen.
Eingaben	• Einblicke in Daten • Werkzeuge für die Erzählung von Geschichten (z. B. Visualisierungssoftware) • Analyse des Publikums
Ausgänge	Datenerzählungen Visuelle Präsentationen Benutzerfreundliche Berichte
Messung des Fortschritts	• Klarheit der Geschichte und Metriken für das Engagement • Rückmeldung der Zielgruppe • Fertigstellung von Visualisierungsmeilensteinen
Allgemeine AI-Risiken	Risiko der Fehlinterpretation oder Übervereinfachung von Daten Unbeabsichtigte Verzerrungen in Erzählungen Ethische Implikationen der Beeinflussung von Entscheidungen durch Geschichtenerzählen

Tabelle 10: Narrative Erkundung

Das Zielpublikum verstehen

Bestimmen Sie die Zielgruppe für die Erzählung, einschließlich der Interessenvertreter, Entscheidungsträger und anderer relevanter Parteien. In diesem Schritt geht es darum, den Wissensstand, die Interessen und Erwartungen der Zielgruppe zu verstehen, um die Erzählung entsprechend anzupassen.

- **Inputs:**
  - Stakeholder-Profile und Kommunikationspräferenzen
  - Projektziele und zu erbringende Leistungen
  - Feedback oder frühere Präsentationen

- **Aktivitäten**:
  - Führen Sie Interviews oder Umfragen durch, um die Präferenzen des Publikums zu ermitteln.
  - Analysieren Sie die Rollen der Stakeholder und die Anforderungen an die Entscheidungsfindung
  - Segmentierung des Publikums auf der Grundlage seiner technischen Kenntnisse und Ziele

- **Outputs:**
  - Dokument zum Publikumsprofil
  - Kommunikationsstrategie
  - Segmentierung der Stakeholder

Definition der Kernbotschaft

Bestimmen Sie die wichtigsten Erkenntnisse oder Schlussfolgerungen, die vermittelt werden sollen. Dieser Schritt stellt sicher, dass die Erzählung eine fokussierte und wirkungsvolle Botschaft enthält, die mit den Projektzielen übereinstimmt.

- **Inputs:**
  - Validierte Datenanalyseergebnisse
  - Unternehmensziele und strategische Prioritäten
  - Vorlieben und Feedback des Publikums

- **Aktivitäten**:
    - Identifizierung der wichtigsten Erkenntnisse und Schlussfolgerungen aus den Daten
    - Priorisierung von Schlüsselbotschaften, die mit den Unternehmenszielen übereinstimmen
    - Verfassen einer Zusammenfassung der Erzählung auf hohem Niveau

- **Outputs:**
    - Dokument mit Schlüsselbotschaften
    - Zusammenfassung der wichtigsten Erkenntnisse
    - Priorisierter Rahmen für die Nachrichtenübermittlung

Gestaltung von Visualisierungs- und Storytelling-Elementen

Entwickeln Sie visuelle Elemente, Diagramme und Erzählhilfen, die die Erzählung ansprechend und leicht verständlich machen. Dazu gehört auch die Auswahl der richtigen Visualisierungstechniken, um Erkenntnisse effektiv zu vermitteln.

- **Inputs:**
    - Dateneinblicke und Analyseergebnisse
    - Visualisierungswerkzeuge (z. B. Tableau, Power BI, Python-Bibliotheken)
    - Bewährte Verfahren für die Erzählung und Visualisierung von Daten

- **Aktivitäten**:
    - Geeignete Visualisierungstechniken auswählen
    - Entwicklung von Prototypen für visuelle Elemente und Einholung von Feedback
    - Kombinieren Sie Visualisierungen mit textlichen oder audiovisuellen Erzähltechniken

- **Outputs:**
    - Visuelle Prototypen
    - Hilfsmittel zum Erzählen von Geschichten (z. B. Infografiken, Folien)
    - Visualisierung von Feedback-Berichten

Ausarbeitung der Erzählung

Erstellen Sie eine fesselnde und strukturierte Geschichte, die die Datenerkenntnisse mit den Geschäftszielen verbindet und dabei visuelle Elemente und Schlüsselbotschaften einbezieht. Die Erzählung sollte logisch, fesselnd und überzeugend sein.

- **Inputs:**
  - Schlüsselbotschaft und priorisierte Erkenntnisse
  - Zielgruppenprofile und Kommunikationspräferenzen
  - Prototypen für Visualisierung und Storytelling
- **Aktivitäten**:
  - Eine strukturierte Erzählung mit klaren Abschnitten schreiben
  - Integrieren Sie Visualisierungen nahtlos in die Geschichte
  - Verwenden Sie Techniken des Geschichtenerzählens, um die Erzählung einprägsam zu machen
- **Outputs:**
  - Entwurf eines erzählenden Dokuments
  - Integrierte visuell-geschichtliche Darstellung
  - Strukturierter Rahmen für das Geschichtenerzählen

Iteration und Verfeinerung der Erzählung

Sammeln Sie Feedback von Stakeholdern oder Testpublikum, um die Klarheit, das Engagement und die Wirksamkeit der Erzählung zu verbessern. Dieser Schritt stellt sicher, dass die Erzählung bei der Zielgruppe ankommt.

- **Inputs:**
  - Entwurf eines erzählenden Dokuments und Visualisierungen
  - Feedback und Vorschläge von Interessengruppen
  - Ergebnisse von Publikumstests (falls zutreffend)
- **Aktivitäten**:
  - Durchführung von Feedback-Sitzungen mit wichtigen Interessengruppen
  - Überarbeitung der Erzählung auf der Grundlage von Feedback und Publikumstests
  - Sicherstellung der Übereinstimmung zwischen der Erzählung und den Unternehmenszielen

- **Outputs:**
    - Verfeinertes erzählerisches Dokument
    - Aktualisierte visuell-geschichtliche Darstellung
    - Zusammenfassung des Stakeholder-Feedbacks

Die Erzählung liefern

Präsentieren Sie den Bericht der Zielgruppe mit dem gewählten Medium, z. B. in Form von Präsentationen, Berichten oder interaktiven Dashboards. Dieser Schritt stellt sicher, dass die Erkenntnisse effektiv vermittelt werden und zur Entscheidungsfindung beitragen.

- **Inputs:**
    - Abgeschlossene Erzählung und visuelle Präsentation der Geschichte
    - Kommunikationsmittel (z. B. Präsentationssoftware, Dashboards)
    - Publikumslogistik (z. B. Sitzungspläne, Plattformpräferenzen

- **Aktivitäten**:
    - Die Erzählung in einer Präsentation, einem Bericht oder einem interaktiven Format wiedergeben
    - Moderation von Diskussionen, um Verständnis und Engagement sicherzustellen
    - Auf Fragen eingehen und bei Bedarf zusätzlichen Kontext liefern

- **Outputs:**
    - Abgegebene erzählende Präsentation
    - Feedback von Interessengruppen
    - Aufgezeichnete Fragen und Antworten

Messung der narrativen Wirkung

Beurteilen Sie die Effektivität des Narrativs in Bezug auf das Erreichen der beabsichtigten Ziele, wie z. B. die Beeinflussung von Entscheidungen, die Einbindung von Stakeholdern oder die Förderung von Aktionen.

- **Inputs:**
    - Erzählungen und Feedback von Interessengruppen
    - Metriken zur Messung des Engagements oder der Auswirkungen (z. B. Änderungen bei Entscheidungen, Grad der Anpassung)
    - Folgeerhebungen oder Befragungen von Interessengruppen

- **Aktivitäten**:
    - Überwachen Sie die Handlungen oder Entscheidungen der Stakeholder, die sich aus der Erzählung ergeben
    - Führen Sie Nacherhebungen oder Interviews durch, um Verständnis und Engagement zu messen
    - Dokumentation der gewonnenen Erkenntnisse zur Verbesserung künftiger Erzählungen

- **Outputs:**
    - Narrativer Wirkungsbericht
    - Metriken zum Engagement
    - Dokumentation der gewonnenen Erkenntnisse

Kategorie	Sondierungstätigkeiten
ID	EX-06
Beschreibung	Hier werden potenzielle Produkte oder Dienstleistungen definiert, die aus den Datenerkenntnissen entstehen können. In dieser Phase werden innovative Anwendungen für die Erkenntnisse erforscht, z. B. die Entwicklung neuer Tools, Anwendungen oder Kundenlösungen. Ziel der Produktexploration ist es, Datenerkenntnisse zu nutzen, um greifbare Produkte oder Dienstleistungen zu schaffen, die einen Mehrwert für Kunden oder Stakeholder darstellen.
Eingaben	• Einblicke in Daten • Fahrplan für die Produktentwicklung • Zielnutzer-Personas • Dokumentation der verfügbaren technischen Mittel
Ausgänge	Produktkonzepte Prototypen Technische Daten Durchführbarkeitsbericht
Messung des Fortschritts	• Entwicklung von Produktkonzepten • Feedback zum Prototyp • Abstimmung der Produktmerkmale auf die Nachfrage • Abschluss von Machbarkeitsstudien und Tests
Allgemeine AI-Risiken	Missbrauch von Produkten Ethische Risiken im Zusammenhang mit Produktanwendungen Sicherheitsbedenken, wenn sensible Daten betroffen sind Potenziell unbeabsichtigte Folgen bei Anwendungen mit Kundenkontakt Herausforderungen bei der Aufrechterhaltung von Transparenz und Vertrauen der Nutzer

Tabelle 11: Produkt-Erkundung

Identifizierung potenzieller Produktkonzepte

Brainstorming von Ideen für Produkte oder Dienstleistungen, die aus datengestützten Erkenntnissen entwickelt werden können. Dazu gehört auch die Erforschung innovativer Anwendungen von Daten zur Schaffung von Mehrwert für Kunden, Stakeholder oder interne Prozesse.

- **Inputs:**
  - Dateneinblicke und ermittelte Chancen
  - Unternehmensziele und Marktanalyse
  - Feedback von Interessengruppen und Endnutzern

- **Aktivitäten**:
  - Durchführung von Brainstorming-Sitzungen mit funktionsübergreifenden Teams
  - Erforschung von Branchentrends und erfolgreichen datengesteuerten Produkten
  - Bewerten Sie, wie vorhandene Datenbestände die Produktinnovation vorantreiben können.

- **Outputs:**
  - Erste Liste von Produktkonzepten
  - Bericht über Innovationschancen
  - Priorisierte Produktideen

Validierung der Produktdurchführbarkeit

Bewertung der technischen und betrieblichen Machbarkeit der vorgeschlagenen Konzepte, einschließlich der Verfügbarkeit von Ressourcen, der technologischen Anforderungen und der Marktnachfrage.

- **Inputs:**
  - Liste der Produktkonzepte
  - Bewertung der Verfügbarkeit von Ressourcen (z. B. Budget, Personal, Infrastruktur)
  - Marktanalyse und Benchmarking der Wettbewerber

- **Aktivitäten**:
    - Bewertung der Datenqualität und Eignung für das vorgeschlagene Produkt
    - Ermittlung der technischen Herausforderungen und des Ressourcenbedarfs
    - Bewertung der Marktnachfrage und Anpassung an die Kundenbedürfnisse

- **Outputs:**
    - Durchführbarkeitsbericht
    - Auswahlliste der realisierbaren Produktideen
    - Zusammenfassung der technischen und Marktvalidierung

Entwurf von Produktprototypen

Entwickeln Sie erste Prototypen oder Proof-of-Concept-Modelle für die in die engere Wahl gekommenen Produktideen. Dieser Schritt hilft dabei, das Produkt zu visualisieren und seine Funktionalität zu testen.

- **Inputs:**
    - In die engere Wahl gekommene Produktideen
    - Design- und Prototyping-Tools (z. B. Wireframing-Software, Entwicklungsplattformen)
    - Technische, funktionale und nicht-funktionale Anforderungen

- **Aktivitäten**:
    - Erstellen von Wireframes oder Mockups zur Darstellung von Produktfunktionen
    - Entwickeln Sie ein Minimum Viable Product (MVP), um die Kernfunktionalitäten zu testen.
    - Sammeln Sie erste Rückmeldungen von Interessengruppen oder Testnutzern

- **Outputs:**
    - Produktprototypen
    - MVPs
    - Feedback-Berichte zu Prototypen

Testen und Iterieren von Prototypen

Testen Sie die Prototypen in kontrollierten Umgebungen oder mit einer Untergruppe von Nutzern, um Feedback zu sammeln und Bereiche mit Verbesserungsbedarf zu ermitteln. Iterationen helfen, das Produkt zu verfeinern, um die Bedürfnisse und Erwartungen der Nutzer besser zu erfüllen.

- **Inputs:**
  - Produktprototypen und MVPs
  - Nutzerfeedback und Bewertungsmetriken testen
  - Tools für Leistungs- und Gebrauchstauglichkeitstests
- **Aktivitäten**:
  - Durchführung von Usability-Tests mit Zielnutzern
  - Analysieren von Leistungskennzahlen aus Prototypentests
  - Umsetzung iterativer Verbesserungen auf der Grundlage von Nutzerfeedback
- **Outputs:**
  - Verbesserte Produktprototypen
  - Berichte über Benutzerfreundlichkeit und Leistung
  - Iterationsprotokolle

Bewertung der Markttauglichkeit

Bewerten Sie, ob das weiterentwickelte Produkt den Marktbedürfnissen und -erwartungen entspricht, und stellen Sie sicher, dass es für die Annahme und den Erfolg positioniert ist.

- **Inputs:**
  - Verfeinerte Prototypen und MVPs
  - Marktforschung und Kundenfeedback
  - Analyse der Wettbewerber

- **Aktivitäten**:
    - Vergleich der Produktmerkmale und des Wertversprechens mit den Marktanforderungen
    - Durchführung von Fokusgruppen oder Umfragen zur Ermittlung des Marktinteresses
    - Analysieren Sie die Wettbewerbsposition und potenzielle Einführungshindernisse

- **Outputs:**
    - Analyse der Marktfähigkeit
    - Feedback zur Kundenvalidierung
    - Rahmen für die wettbewerbsfähige Positionierung

Entwicklung einer Go-to-Market-Strategie

Erstellen Sie einen detaillierten Plan für die Markteinführung des Produkts, einschließlich Marketing-, Vertriebs- und Adoptionsstrategien. Dieser Schritt stellt sicher, dass das Produkt die richtige Zielgruppe effektiv erreicht.

- **Inputs:**
    - Marktanpassungsanalyse und Kundenfeedback
    - Organisatorische Marketing- und Vertriebsressourcen
    - Verkaufs- und Preisstrategien

- **Aktivitäten**:
    - Definition von Zielkundensegmenten und Marketingbotschaften
    - Vertriebskanäle und Partnerschaften planen
    - Festlegung von Preismodellen und Anreizen für die Einführung

- **Outputs:**
    - Dokument zur Markteinführungsstrategie
    - Marketingplan
    - Rahmen für Preisgestaltung und Vertrieb

Markteinführung und Überwachung des Produkts

Einführung des Produkts auf dem vorgesehenen Markt und kontinuierliche Überwachung seiner Leistung, um Verbesserungsmöglichkeiten zu ermitteln und sicherzustellen, dass es die angestrebten Ziele erreicht.

- **Inputs:**
  - Fertigstellung der Produkt- und Markteinführungsstrategie
  - Analysetools zur Verfolgung der Produktleistung
  - Kundenfeedback und Supportkanäle
- **Aktivitäten**:
  - Durchführung der Produkteinführungskampagne
  - Überwachen Sie wichtige Leistungsindikatoren (KPIs) wie Nutzerakzeptanz, Kundenbindung und Umsatz
  - Einholen von Feedback nach der Einführung und Behebung etwaiger Probleme
- **Outputs:**
  - Bericht über die Produkteinführung
  - Dashboard zur KPI-Überwachung
  - Verbesserungsplan für die Zeit nach der Markteinführung

Das Produkt erhalten und weiterentwickeln

Stellen Sie sicher, dass das Produkt im Laufe der Zeit relevant und wettbewerbsfähig bleibt, indem Sie Updates, neue Funktionen oder Verbesserungen auf der Grundlage von Benutzerfeedback und Markttrends einführen.

- **Inputs:**
  - Feedback und KPI-Berichte nach der Einführung
  - Analyse der Markttrends und Aktualisierung der Wettbewerber
  - Organisatorische Ressourcen für die laufende Entwicklung

- **Aktivitäten**:
    - Entwicklung einer Roadmap für Produktaktualisierungen
    - Einführung neuer Funktionen oder Verbesserungen auf der Grundlage der Bedürfnisse der Nutzer
    - Beobachtung von Markttrends und Anpassung der Produktstrategien nach Bedarf

- **Outputs:**
    - Fahrplan für die Produktaktualisierung
    - Berichte über Funktionserweiterungen
    - Langfristige Produktstrategie

Zielgerichtete Komponenten (CRISP-DM)

Geschäftsverständnis

Kategorie	Zielgerichtete Aktivitäten
ID	GO-01
Beschreibung	Hier geht es darum, ein klares Verständnis des geschäftlichen Kontextes, der Ziele und der Anforderungen des Projekts zu entwickeln. In dieser Phase wird sichergestellt, dass die Data-Science-Initiative mit den strategischen Zielen des Unternehmens in Einklang steht. Sie umfasst das Sammeln von Erkenntnissen über potenzielle Einschränkungen, betriebliche Anforderungen und wichtige Erfolgskriterien sowie die Definition messbarer Ziele zur Bewertung der Projektergebnisse.
Eingaben	• Detaillierte geschäftliche Anforderungen • Befragung von Interessenvertretern • Organisatorische Ziele • Analyse der Wettbewerber • Risikobewertung
Ausgänge	Projekt-Charta Zusammenfassung der Unternehmensziele Dokument zum Projektumfang Bericht über die Anpassung an die Interessengruppen Erfolgskriterien
Messung des Fortschritts	• Fertigstellung einer Projektcharta • Klare Formulierung von Geschäftszielen und -beschränkungen • Zustimmung und Abstimmung mit den Interessengruppen • Verfolgung von Meilensteinen bei vereinbarten Zielen
Allgemeine AI-Risiken	Diskrepanz zwischen Projektergebnissen und Geschäftsanforderungen Unzureichende Einbeziehung von Interessengruppen, was zu unklaren Zielen führt Potenzielle ethische oder rechtliche Fragen, die in den ursprünglichen Anforderungen übersehen wurden

Tabelle 12: Business-Verständnis

Definieren des Geschäftsproblems

Formulieren Sie klar und deutlich das Geschäftsproblem, das mit Hilfe von Data Science gelöst werden soll. Dieser Schritt gewährleistet die Ausrichtung an den Unternehmenszielen und bildet die Grundlage für die nachfolgenden Aktivitäten.

- **Inputs:**
  - Befragung von Interessenvertretern und Feedback
  - Organisatorische Ziele und KPIs
  - Branchen-Benchmarks und Wettbewerbsanalysen

- **Aktivitäten**:
  - Durchführung von Workshops oder Interviews mit Interessenvertretern zur Ermittlung von Herausforderungen
  - Analyse der organisatorischen Ziele zur Sicherstellung der Problemabstimmung
  - Formulierung des Problems in Form von spezifischen, messbaren Zielen

- **Outputs:**
  - Dokument zur Problemdefinition
  - Bericht zur Ausrichtung der Unternehmensziele
  - Erste Erfolgskriterien

Verstehen des geschäftlichen Kontextes

Gewinnen Sie ein umfassendes Verständnis des Geschäftsumfelds, einschließlich betrieblicher Arbeitsabläufe, Markttrends und Kundenverhalten. Dieser Schritt hilft bei der Kontextualisierung des Problems und der Verfeinerung der Projektziele.

- **Inputs:**
  - Branchenberichte und Marktanalysen
  - Kundenfeedback und Verhaltensdaten
  - Dokumentation der internen Prozesse und Arbeitsabläufe

- **Aktivitäten**:
    - Überprüfung von Branchentrends und Kundenpräferenzen
    - Abbildung der für das Problem relevanten organisatorischen Prozesse und Arbeitsabläufe
    - Identifizierung externer Faktoren (z. B. wirtschaftliche, regulatorische), die sich auf das Unternehmen auswirken

- **Outputs:**
    - Bericht über den geschäftlichen Kontext
    - Prozess-Karten
    - Externe Faktorenanalyse

Zielsetzung und Erfolgskriterien festlegen

Definieren Sie klare Ziele und Erfolgskriterien, die sich an den Unternehmenszielen orientieren. Diese Kriterien dienen als Benchmarks, um den Projekterfolg zu messen.

- **Inputs:**
    - Dokument zur Problemdefinition
    - Organisatorische KPIs und strategische Prioritäten
    - Beiträge der Interessengruppen zu den gewünschten Ergebnissen

- **Aktivitäten**:
    - Entwicklung spezifischer, messbarer, erreichbarer, relevanter und zeitgebundener (SMART) Ziele
    - Definition von qualitativen und quantitativen Erfolgskriterien
    - Abgleich der Erfolgskriterien mit den Erwartungen der Stakeholder

- **Outputs:**
    - Dokument mit SMART-Zielen
    - Bericht über die Erfolgskriterien
    - Vereinbarung zur Anpassung an die Interessengruppen

Scoping des Projekts

Definieren Sie den Umfang des Data-Science-Projekts, einschließlich der zu erbringenden Leistungen, des Zeitplans, der Ressourcen und der Beschränkungen. Dieser Schritt verhindert eine schleichende Ausweitung des Umfangs und sorgt für Klarheit darüber, was machbar ist.

- **Inputs:**
  - Problemdefinition und Unternehmensziele
  - Bewertung der organisatorischen Ressourcen und Kapazitäten
  - Erwartungen der Stakeholder und Zeitvorgaben
- **Aktivitäten**:
  - Ermittlung der wichtigsten Ergebnisse und Meilensteine
  - Erstellen eines übergeordneten Projektzeitplans und Zuweisung von Ressourcen
  - Dokumentieren Sie potenzielle Risiken und Beschränkungen
- **Outputs:**
  - Dokument zum Projektumfang
  - Zeitplan für Meilensteine
  - Bericht zur Risikobewertung

Identifizierung der wichtigsten Stakeholder

Bestimmen Sie die Personen oder Gruppen, die ein persönliches Interesse am Erfolg des Projekts haben, und legen Sie deren Rollen und Zuständigkeiten fest. Dieser Schritt gewährleistet eine effektive Kommunikation und Zusammenarbeit.

- **Inputs:**
  - Organisatorische Struktur und Teamhierarchien
  - Projektziele und Umfang
  - Befragungen von Interessengruppen und Rollenbewertungen
- **Aktivitäten**:
  - Erstellen einer Stakeholder-Map zur Ermittlung der wichtigsten Akteure
  - Zuweisung von Rollen und Verantwortlichkeiten an die Beteiligten
  - Entwicklung eines Kommunikationsplans, um die Beteiligten auf dem Laufenden zu halten

- **Outputs:**
    - Karte der Interessengruppen
    - Dokument über Rollen und Verantwortlichkeiten
    - Kommunikationsplan

Definition von Beschränkungen und Risiken

Identifizieren Sie potenzielle Einschränkungen und Risiken, die sich auf das Projekt auswirken könnten, wie z. B. Ressourcenbeschränkungen, Einhaltung von Vorschriften und ethische Überlegungen.

- **Inputs:**
    - Umfang und Ziele des Projekts
    - Organisatorische Maßnahmen und Risikoregister
    - Regulatorische und Compliance-Anforderungen

- **Aktivitäten**:
    - Ermittlung von Ressourcenbeschränkungen (z. B. Budget, Zeit, Fachwissen)
    - Bewertung der rechtlichen und ethischen Aspekte der Datennutzung
    - Dokumentation potenzieller Risiken und Entwicklung von Strategien zur Risikominderung

- **Outputs:**
    - Bericht über Risiken und Sachzwänge
    - Plan zur Risikominderung
    - Checkliste zur Einhaltung der Vorschriften

Dokumentieren und Kommunizieren des Geschäftsverständnisses

Fassen Sie alle Ergebnisse, Ziele und Einschränkungen in einem formellen Dokument zusammen, um die Abstimmung zwischen den Beteiligten zu gewährleisten. Dieser Schritt fördert die Transparenz und das gemeinsame Verständnis.

- **Inputs:**
    - Problemstellung und Ziele
    - Dokumente zur Anpassung an die Interessengruppen
    - Analyse der Risiken und Beschränkungen

- **Aktivitäten**:
    - Erstellen eines Berichts zum Geschäftsverständnis, der alle Teilaufgaben konsolidiert
    - Entwicklung von Präsentationsmaterialien für Interessengruppen
    - Veranstalten Sie ein Kickoff-Meeting, um die Ergebnisse und den Projektplan zu kommunizieren.

- **Outputs:**
    - Bericht zum Geschäftsverständnis
    - Materialien für die Präsentation von Interessengruppen
    - Protokoll des Kickoff-Meetings

Kategorie	Zielgerichtete Aktivitäten
ID	GO-02
Beschreibung	Konzentriert sich darauf, ein umfassendes Verständnis der für das Projekt relevanten Daten zu erlangen. Diese Phase umfasst die Bewertung von Datenquellen, die Analyse der Datenqualität, die Ermittlung von Lücken und das Verständnis der Beziehungen innerhalb der Daten. Sie hilft festzustellen, ob die Daten die Geschäftsziele angemessen erfüllen können.
Eingaben	• Rohdaten aus verschiedenen Quellen • Datenschema und Struktur • Metriken zur Datenqualität • Metadaten • Erste Werkzeuge zur Datenexploration
Ausgänge	Datenwörterbuch Bericht über die Datenqualität Bewertung der Vollständigkeit der Daten Erste Ergebnisse der explorativen Datenanalyse (EDA)
Messung des Fortschritts	• Abschluss der Datenprofilerstellung und Qualitätsbewertung • Identifizierung der wichtigsten Datenmerkmale und potenzieller Probleme • Dokumentation von Datenlücken oder Einschränkungen
Allgemeine AI-Risiken	Schlechte Datenqualität was zu ungenauen oder verzerrten Ergebnissen führt Unzureichende Daten für das Modelltraining Überanpassung an bestimmte Datenmerkmale, wenn diese nicht repräsentativ für reale Szenarien sind

Tabelle 13: Verständnis der Daten

Identifizierung von Datenquellen

Identifizieren Sie alle potenziellen Datenquellen, die für das Projekt relevant sein könnten. Dieser Schritt konzentriert sich auf die Katalogisierung interner und externer Datensätze, um eine umfassende Abdeckung des Problembereichs zu gewährleisten.

- **Inputs:**
  - Unternehmensziele und Problemdefinition
  - Organisatorischer Datenbestand
  - Externe Daten-Repositories, APIs oder Drittanbieter
- **Aktivitäten**:
  - Durchführung einer Bestandsaufnahme der vorhandenen Datenbestände des Unternehmens
  - Recherche externer Datensätze und Datenanbieter
  - Dokument-Metadaten, einschließlich Zuverlässigkeit der Quelle und Zugänglichkeit
- **Outputs:**
  - Katalog der potenziellen Datenquellen
  - Bericht über Metadaten
  - Zusammenfassung der Datenzugänglichkeit

Sammeln erster Datenproben

Sammeln Sie erste Stichproben von identifizierten Datenquellen, um deren Struktur, Qualität und potenzielle Relevanz für das Projekt zu verstehen.

- **Inputs:**
  - Katalog der potenziellen Datenquellen
  - Zugangsdaten und Berechtigungen für den Datenabruf
  - Datenextraktionswerkzeuge oder Pipelines (z. B. APIs, ETL-Systeme)

- **Aktivitäten**:
    - Abrufen von Beispieldatensätzen oder Abfragen von Datenbeständen
    - Extrahieren und Speichern von Datenproben für eine erste Analyse
    - Sicherstellung der Einhaltung von Zugangsprotokollen und Datenschutzrichtlinien

- **Outputs:**
    - Gesammelte Datenproben
    - Muster eines Lagerberichts
    - Zugang zu Compliance-Protokollen

Durchführen einer explorativen Datenanalyse (EDA)

Führen Sie eine vorläufige Analyse durch, um die Struktur, Verteilungen, Beziehungen und Anomalien der Daten zu untersuchen. Dieser Schritt hilft dabei, Muster und potenzielle Probleme bereits in einem frühen Stadium des Projekts zu erkennen.

- **Inputs:**
    - Erste Datenproben
    - Datenanalysetools (z. B. Python, R, Tableau, Excel)
    - Fachwissen über den Kontext der Daten

- **Aktivitäten**:
    - Erstellen von deskriptiven Statistiken über die zugrunde liegenden Datenverteilungen
    - Erstellen Sie Visualisierungen, um Beziehungen und Anomalien aufzuzeigen.
    - Identifizieren Sie fehlende Werte, Ausreißer und potenzielle Verzerrungen in den Daten.

- **Outputs:**
    - EDA-Bericht
    - Dashboards zur Datenvisualisierung
    - Erste Erkenntnisse und Anomalien protokollieren

Bewertung der Datenqualität

Bewerten Sie die Qualität der Daten anhand von Kriterien wie Genauigkeit, Vollständigkeit, Konsistenz, Aktualität und Zuverlässigkeit. Dieser Schritt stellt sicher, dass die Daten für den vorgesehenen Zweck geeignet sind.

- **Inputs:**
  - Datenbeispiele und EDA-Ergebnisse
  - Kriterien und Instrumente zur Bewertung der Datenqualität
  - Metadaten zur Beschreibung der Datenquellen
- **Aktivitäten**:
  - Prüfung auf fehlende oder unvollständige Aufzeichnungen
  - Bewertung der Konsistenz von Datensätzen oder Attributen
  - Validierung der Datengenauigkeit anhand bekannter Benchmarks oder Standards
- **Outputs:**
  - Bericht zur Bewertung der Datenqualität
  - Liste der Datenqualitätsprobleme
  - Empfehlungen zur Datenbereinigung

Identifizierung von Datenbeziehungen und Abhängigkeiten

Analysieren Sie die Beziehungen und Abhängigkeiten zwischen den Variablen im Datensatz, um mögliche Kausalzusammenhänge aufzudecken.

- **Inputs:**
  - EDA-Ergebnisse und Visualisierungen
  - Fachwissen und Sachkenntnis
  - Datenwörterbücher oder Schemadokumentation
- **Aktivitäten**:
  - Korrelations- und Kovarianzanalyse durchführen
  - Clustering- oder Klassifizierungsverfahren verwenden, um ähnliche Datenpunkte zu gruppieren
  - Identifizierung von Schlüsselvariablen und deren Abhängigkeiten

- **Outputs:**
    - Bericht zur Beziehungsanalyse
    - Rangfolge der Wichtigkeit der Variablen
    - Abhängigkeitskarten

Verstehen von Datenbeschränkungen und Risiken

Identifizieren Sie Einschränkungen, Verzerrungen oder Risiken im Zusammenhang mit den Daten. Dieser Schritt stellt sicher, dass das Bewusstsein für potenzielle Herausforderungen geschärft wird und hilft, Probleme in späteren Phasen zu entschärfen.

- **Inputs:**
    - Bericht zur Bewertung der Datenqualität
    - Metadaten und bereichsspezifische Erkenntnisse
    - Feedback der Interessengruppen zur Datenrelevanz

- **Aktivitäten**:
    - Dokumentieren Sie bekannte Verzerrungen oder Lücken in den Daten
    - Risiken des Datenmissbrauchs oder der Dateninterpretation aufzeigen
    - Empfehlen Sie bei Bedarf zusätzliche Datenerhebungen

- **Outputs:**
    - Datenbeschränkungen und Risikobericht
    - Minderungsstrategien (falls zutreffend)
    - Empfehlungen für zusätzliche Datenerhebungen (falls zutreffend)

Dokumentieren und Kommunizieren des Datenverständnisses

Fassen Sie die Ergebnisse, Datenqualitätsbewertungen und Beziehungen in einem umfassenden Dokument zusammen. Vermitteln Sie diese Erkenntnisse den Beteiligten, um ein gemeinsames Verständnis zu gewährleisten.

- **Inputs:**
    - EDA-Bericht und Bewertung der Datenqualität
    - Ergebnisse der Beziehungsanalyse
    - Dokumentation der Risiken und Beschränkungen

- **Aktivitäten**:
    - Erstellung eines umfassenden Berichts zum Datenverständnis
    - Entwicklung von Visualisierungen zur Präsentation wichtiger Ergebnisse
    - Veranstaltung von Stakeholder-Sitzungen, um Datenerkenntnisse zu diskutieren und die nächsten Schritte abzustimmen

- **Outputs:**
    - Bericht zum Datenverständnis
    - Materialien zur Präsentation
    - Zusammenfassung des Stakeholder-Feedbacks

Kategorie	Zielgerichtete Aktivitäten
ID	GO-03
Beschreibung	Umfasst die Bereinigung, Umwandlung und Organisation der Daten in ein für die Modellierung geeignetes Format. Dieser Schritt umfasst den Umgang mit fehlenden Werten, die Kodierung kategorischer Variablen und die Normalisierung der Daten. Eine effektive Datenaufbereitung stellt sicher, dass die in die Modelle eingespeisten Daten genau, konsistent und frei von irrelevantem Rauschen sind.
Eingaben	• Rohdaten • Datenbereinigungstools (z. B. ETL-Tools) • Umwandlungs- und Normalisierungsalgorithmen • Ressourcen zur Datenerweiterung
Ausgänge	Vorverarbeiteter Datensatz Transformierte Daten bereit für die Modellierung Merkmalsgestützte Daten Datendokumentation und Transformationsprotokoll
Messung des Fortschritts	• Abschluss der Datenbereinigung • Validierung anhand von Qualitätsstandards • Dokumentation der Vorverarbeitungsschritte • Überprüfung der aufbereiteten Daten durch die Interessengruppen
Allgemeine AI-Risiken	Datenlecks durch zukünftige Informationen in den Trainingsdaten Überanpassung durch übermäßiges Feature Engineering Versehentliches Entfernen wichtiger Informationen während der Reinigung

Tabelle 14: Vorbereitung der Daten

Auswählen relevanter Daten

Identifizieren und extrahieren Sie die Teilmengen von Daten, die für die Projektziele relevant sind, und stellen Sie sicher, dass nur notwendige und nützliche Daten in die nachfolgenden Schritte aufgenommen werden.

- **Inputs:**
  - Bericht zum Datenverständnis
  - Geschäftsziele und Anforderungen an Anwendungsfälle
  - Metadaten oder Schema der verfügbaren Datensätze
- **Aktivitäten**:
  - Filtern von Datenattributen auf der Grundlage von Projektanforderungen
  - Redundante oder irrelevante Spalten/Zeilen entfernen
  - Dokumentieren Sie die Auswahlkriterien und die Gründe für die Ausschlüsse
- **Outputs:**
  - Ausgewählte Teilmenge der relevanten Daten
  - Protokoll der Datenselektion
  - Dokumentation der Kriterien

Daten reinigen

Behebung von Problemen mit der Datenqualität, wie fehlende Werte, Ausreißer und Inkonsistenzen, um sicherzustellen, dass der Datensatz zuverlässig und für die Analyse und Modellierung geeignet ist.

- **Inputs:**
  - Teilmenge der Daten auswählen
  - Bericht zur Bewertung der Datenqualität
  - Reinigungswerkzeuge (z. B. Python-Bibliotheken wie Pandas, R-Pakete)
- **Aktivitäten**:
  - Umgang mit fehlenden Werten durch Imputation, Löschung oder Substitution
  - Entfernen oder Anpassen von Ausreißern auf der Grundlage statistischer Schwellenwerte
  - Standardisierung und Korrektur inkonsistenter Dateneingaben

- **Outputs:**
    - Bereinigter Datensatz
    - Bericht über die Reinigungsmethodik
    - Aktualisiertes Protokoll der Qualitätsbewertung

Daten umwandeln

Wenden Sie Transformationen an, um die Daten zu standardisieren und sie mit Modellierungs- oder Analysetechniken kompatibel zu machen. Dazu gehören Normalisierung, Kodierung und Skalierung.

- **Inputs:**
    - Bereinigter Datensatz
    - Anforderungen an die Modellierung oder Analyse
    - Techniken der Umwandlung

- **Aktivitäten**:
    - Normalisieren Sie numerische Daten, um die Konsistenz der Attribute zu gewährleisten.
    - Kodierung kategorischer Variablen in numerische Formate
    - Datenattribute auf vergleichbare Bereiche skalieren

- **Outputs:**
    - Transformierter Datensatz
    - Protokoll der Transformationsmethodik
    - Bericht über die Kompatibilität von Merkmalen

Technische Merkmale

Erstellen Sie neue Funktionen oder ändern Sie bestehende, um die Vorhersagefähigkeiten des Datensatzes zu verbessern und ihn auf die Projektziele abzustimmen.

- **Inputs:**
    - Transformierter Datensatz
    - Fachwissen und Geschäftskenntnisse
    - Tools und Frameworks für das Feature Engineering

- **Aktivitäten**:
    - Kombinieren Sie Attribute, um neue zusammengesetzte Features zu erstellen
    - Ableitung zeitlicher Merkmale (z. B. Saisonalität, Trends)
    - Auswahl oder Entfernung von Merkmalen auf der Grundlage ihrer Relevanz für die Projektziele

- **Outputs:**
    - Feature-engineerter Datensatz
    - Rangfolge der Wichtigkeit von Merkmalen
    - Feature-Engineering-Bericht

Integration mehrerer Datensätze

Kombinieren Sie Daten aus verschiedenen Quellen oder Datensätzen, um einen einheitlichen Datensatz für die Analyse oder Modellierung zu erstellen. Dazu gehört auch die Behandlung von Inkonsistenzen und die Gewährleistung einer nahtlosen Integration.

- **Inputs:**
    - Datensätze aus mehreren Quellen.
    - Schemaabbildung oder Metadaten für jeden Datensatz.
    - Integrationswerkzeuge (z. B. ETL-Pipelines).

- **Aktivitäten**:
    - Schemaabgleich durchführen und Attributbenennungskonflikte auflösen.
    - Zusammenführen von Datensätzen anhand von Schlüsseln (z. B. IDs, Zeitstempel).
    - Beseitigung von Diskrepanzen zwischen den Datenformaten oder Maßstäben der verschiedenen Quellen.

- **Outputs:**
    - Einheitlicher Datensatz
    - Dokumentation der Integrationsmethodik
    - Bericht über die Datenkonsistenz

Datenreduzierung

Vereinfachen Sie den Datensatz, indem Sie ihn verkleinern und dabei die wesentlichen Informationen beibehalten. Dieser Schritt sorgt für Effizienz bei der anschließenden Modellierung und Analyse.

- **Inputs:**
  - Einheitlicher Datensatz
  - Rangfolge der Wichtigkeit von Merkmalen
  - Techniken zur Dimensionalitätsreduzierung
- **Aktivitäten**:
  - Anwendung von Techniken zur Verringerung der Dimensionalität, um die Anzahl der Merkmale zu reduzieren
  - Daten aggregieren, wenn keine Granularität erforderlich ist
  - Eliminieren Sie Attribute mit geringer Varianz oder Relevanz
- **Outputs:**
  - Reduzierter Datensatz
  - Bericht zur Dimensionalitätsreduktion
  - Analyse der Auswirkungen auf die Leistung

Dokumentieren und Kommunizieren der Datenaufbereitung

Fassen Sie alle Schritte der Datenaufbereitung zusammen und stellen Sie sicher, dass die Beteiligten die an den Daten vorgenommenen Änderungen verstehen. Dieser Schritt schafft Transparenz und eine Referenz für zukünftige Iterationen.

- **Inputs:**
  - Bereinigter, transformierter und vorbereiteter Datensatz
  - Protokolle und Berichte über alle Schritte der Datenaufbereitung
  - Anforderungen der Interessengruppen an die Dokumentation

- **Aktivitäten**:
    - Erstellung eines umfassenden Datenaufbereitungsberichts, in dem alle Änderungen detailliert aufgeführt sind
    - Entwicklung von Visualisierungen oder Diagrammen zur Veranschaulichung der wichtigsten Änderungen
    - Weitergabe der Dokumentation an die Beteiligten und Einholung der Genehmigung

- **Outputs:**
    - Bericht zur Datenaufbereitung
    - Materialien für die Kommunikation mit Interessengruppen
    - Zulassungsunterlagen

Kategorie	Zielgerichtete Aktivitäten
ID	GO-04
Beschreibung	Der Prozess der Auswahl und Anwendung geeigneter maschineller Lern- oder statistischer Modelle auf die aufbereiteten Daten, um Vorhersagen oder Erkenntnisse zu gewinnen. Dieser Schritt umfasst die Auswahl von Algorithmen, die Abstimmung von Hyperparametern und die Kreuzvalidierung zur Gewährleistung der Robustheit. Ziel ist es, ein Modell zu erstellen, das die Daten genau wiedergibt und gleichzeitig gut auf neue Daten verallgemeinert werden kann.
Eingaben	• Vorbereitete Daten • Kriterien für die Modellauswahl • Algorithmus-Bibliotheken • Ressourcen für die Datenverarbeitung • Kreuzvalidierungstechniken
Ausgänge	Trainiertes Modell Leistungsmetriken für Modelle Ergebnisse der Kreuzvalidierung Parameter-Einstellungen Modellartefakte (z. B. Modellgewichte)
Messung des Fortschritts	• Erreichen von Leistungszielen • Abschluss der Kreuzvalidierung • Konvergenz der optimalen Modellparameter • Dokumentation der Modellabstimmung und -validierung
Allgemeine AI-Risiken	Überanpassung oder Unteranpassung Abhängigkeit von Modellannahmen Mangelnde Transparenz bei komplexen Modellen, wie z. B. Deep Learning, was zu Problemen bei der Erklärbarkeit führt

Tabelle 15: Modellierung

Auswahl der Modellierungstechniken

Identifizierung geeigneter Modellierungstechniken und -algorithmen auf der Grundlage des Problemtyps, der Datenmerkmale und der Projektziele. Dieser Schritt gewährleistet die Auswahl von Methoden, die mit den geschäftlichen Anforderungen und Datenbeschränkungen übereinstimmen.

- **Inputs:**
  - Vorbereiteter Datensatz
  - Projektziele und Erfolgskriterien
  - Verfügbare Algorithmen und Modellierungsrahmen
- **Aktivitäten**:
  - Klassifizierung des Problemtyps (z. B. überwacht, unbeaufsichtigt)
  - Recherche und Vorauswahl geeigneter Algorithmen oder Modellierungsansätze
  - Abwägen von Kompromissen wie Interpretierbarkeit, Skalierbarkeit und Leistung
- **Outputs:**
  - Auswahlliste der Modellierungstechniken
  - Begründungsdokument für ausgewählte Techniken
  - Modellierungsplan

Aufteilung der Daten

Unterteilen Sie den Datensatz in Trainings-, Validierungs- und Testuntermengen, um eine angemessene Modellentwicklung, -abstimmung und -bewertung zu ermöglichen.

- **Inputs:**
  - Vorbereiteter Datensatz
  - Leitlinien für die Aufteilung der Daten (z. B. 70-20-10 oder Kreuzvalidierungsfalten)
- **Aktivitäten**:
  - Aufteilung des Datensatzes in Trainings-, Validierungs- und Testsätze
  - Sicherstellung des Klassengleichgewichts bei Splits (bei Klassifizierungsproblemen)
  - Dokumentieren Sie die Methodik der Datenteilung und des Datenabgleichs

- **Outputs:**
    - Trainings-, Validierungs- und Testdatensätze
    - Datenaufteilung und Ausgleichsprotokolle

Erste Modelle entwickeln

Trainieren Sie die Modelle mit den ausgewählten Techniken anhand der Trainingsdaten, um erste Ergebnisse zu erzielen. Dieser Schritt liefert eine Grundlage für die weitere Abstimmung und Bewertung.

- **Inputs:**
    - Trainingsdatensatz
    - Ausgewählte Modellierungstechniken
    - Modellierungswerkzeuge und Bibliotheken

- **Aktivitäten**:
    - Modelle mit Standard- oder Basiskonfiguration trainieren
    - Dokumentation der Ausgangsleistung für jedes Modell
    - Erste Ergebnisse vergleichen, um Stärken und Schwächen zu erkennen

- **Outputs:**
    - Erste Modelle
    - Grundlegende Leistungsmetriken
    - Ausbildungsprotokolle

Abstimmung der Modellparameter

Optimierung der Hyperparameter zur Verbesserung der Modellleistung. Dieser Schritt umfasst systematische Anpassungen, um die beste Konfiguration für das gewählte Modell zu finden.

- **Inputs:**
    - Erste Modelle und Basiskennzahlen
    - Methoden zur Abstimmung der Hyperparameter (z. B. Gittersuche, Bayes'sche Optimierung)
    - Validierungsdatensatz

- **Aktivitäten**:
    - Identifizierung von Schlüsselhyperparametern für die Abstimmung
    - Hyperparameter-Optimierung anhand des Validierungsdatensatzes durchführen
    - Aufzeichnung der Ergebnisse und Ermittlung der leistungsstärksten Konfigurationen

- **Outputs:**
    - Optimierte Modelle
    - Hyperparameter-Abstimmungsbericht
    - Aktualisierte Leistungsmetriken

Bewertung der Modellleistung

Bewerten Sie die Leistung der Modelle anhand von Validierungs- und Testdatensätzen. Dieser Schritt gewährleistet, dass die ausgewählten Modelle gut verallgemeinert werden können und die Erfolgskriterien erfüllen.

- **Inputs:**
    - Optimierte Modelle
    - Validierungs- und Testdatensätze
    - Leistungsmetriken

- **Aktivitäten**:
    - Bewertung der Modellleistung mit Hilfe geeigneter Metriken
    - Vergleich von Modellen auf der Grundlage ihrer Leistung und Eignung
    - Identifizierung des leistungsfähigsten Modells für den Einsatz

- **Outputs:**
    - Bericht über die Bewertung des Modells
    - Vergleichende Leistungsanalyse
    - Ausgewähltes Modell für den Einsatz

Interpretation und Validierung des Modells

Sicherstellen, dass das ausgewählte Modell interpretierbar und zuverlässig ist und den Erwartungen der Interessengruppen entspricht. Dieser Schritt beinhaltet die Validierung des Modells anhand der Geschäftsziele und des Fachwissens.

- **Inputs:**
  - Ausgewählte Modell- und Bewertungsmetriken
  - Fachwissen und Feedback von Interessengruppen
  - Unternehmensziel und Zwänge

- **Aktivitäten**:
  - Analysieren Sie die Vorhersagen des Modells auf Interpretierbarkeit und Relevanz
  - Validierung des Modells anhand der domänenspezifischen Erwartungen
  - Präsentation der Ergebnisse auf der Grundlage von Feedback und Validierung durch die Beteiligten

- **Outputs:**
  - Bericht über die Auslegung des Modells
  - Zusammenfassung des Stakeholder-Feedbacks
  - Bestätigung der Validierung

Dokumentieren und Kommunizieren der Modellentwicklung

Zusammenfassung des Modellierungsprozesses, einschließlich der verwendeten Techniken, der Leistungsergebnisse und der Gründe für die Modellauswahl. Kommunikation der Ergebnisse und Empfehlungen an die Beteiligten.

- **Inputs:**
  - Modellierung von Protokollen und Berichten
  - Bewertungsmaßstäbe und Interpretationsergebnisse
  - Anforderungen an die Kommunikation mit den Stakeholdern

- **Aktivitäten**:
    - Erstellung eines umfassenden Modellierungsberichts, der alle Schritte zusammenfasst
    - Entwicklung von Visualisierungen oder Präsentationen zur Erläuterung der Modellleistung
    - Veranstaltung einer Besprechung zur Abstimmung über das ausgewählte Modell und die nächsten Schritte

- **Outputs:**
    - Modellierungsbericht
    - Materialien für die Präsentation von Interessengruppen
    - Genehmigungsunterlagen für das ausgewählte Modell

Bewertung

Kategorie	Zielgerichtete Aktivitäten
ID	GO-05
Beschreibung	Konzentriert sich auf die Bewertung der Modellleistung in Bezug auf die Unternehmensziele. Dazu gehört die Bewertung der Modellgenauigkeit, Sensitivität und Robustheit. Dazu gehört auch die Überprüfung auf unbeabsichtigte Verzerrungen und die Sicherstellung, dass das Modell die vordefinierten Kriterien für den Einsatz erfüllt.
Eingaben	• Trainiertes Modell • Bewertungsmetriken • Validierungsdatensatz • Metriken zur Unternehmensleistung
Ausgänge	Bericht über die Bewertung des Modells Leistungsvergleich (z. B. Genauigkeit, Präzision, Wiedererkennung) Folgenabschätzung für Unternehmen Bericht zur Fehleranalyse
Messung des Fortschritts	• Erreichen von Leistungsmaßstäben • Überprüfung der Bewertungsergebnisse durch die Interessengruppen • Genehmigung des Modells für den Einsatz oder die weitere Abstimmung • Dokumentation der Grenzen für Business Drift und Data Drift
Allgemeine AI-Risiken	Voreingenommenheit bei der Bewertung durch übermäßiges Vertrauen in bestimmte Messgrößen Verschlechterung der Modellleistung in realen Umgebungen Unzureichende Validierung führt zu Modelldrift

Tabelle 16: Bewertung

Festlegung von Bewertungsmetriken

Wählen Sie Metriken, die mit den Geschäftszielen übereinstimmen und die Leistung des Modells genau bewerten. Die Metriken sollten die spezifischen Ziele des Projekts widerspiegeln, wie z. B. Genauigkeit, Präzision, Rückruf oder RMSE.

- **Inputs:**
  - Unternehmensziel und Erfolgskriterien
  - Ausgewähltes Modell und Problemtyp (z. B. überwacht, unbeaufsichtigt)
  - Metrische Definitionen und Branchen-Benchmarks
- **Aktivitäten**:
  - Ermittlung geeigneter Bewertungsmaßstäbe auf der Grundlage des Problemtyps
  - Definition von Schwellenwerten oder Benchmarks für den Erfolg
  - Validierung der Metrikauswahl mit den Beteiligten
- **Outputs:**
  - Definitionen von Bewertungskennzahlen
  - Dokument über die Erfolgsschwelle
  - Von den Interessengruppen genehmigte Metrikliste

Bewertung der Modellleistung

Messen Sie die Leistung des Modells in den Validierungs- und Testdatensätzen anhand der festgelegten Metriken. In diesem Schritt wird die Fähigkeit des Modells zur Generalisierung auf neue Daten bewertet.

- **Inputs:**
  - Trainierte(s) Modell(e)
  - Validierungs- und Testdatensätze
  - Definierte Bewertungsmetriken
- **Aktivitäten**:
  - Anwendung des Modells auf die Validierungs- und Testdatensätze
  - Berechnung von Leistungskennzahlen
  - Vergleichen Sie die Metriken mit den Erfolgsschwellenwerten

- **Outputs:**
    - Bericht über die Leistungskennzahlen
    - Zusammenfassung der Testergebnisse
    - Vergleich der Validierungsergebnisse

Analyse von Voreingenommenheit und Fairness

Bewertung des Modells im Hinblick auf mögliche Verzerrungen oder unfaire Ergebnisse, die bestimmte Gruppen oder Merkmale betreffen könnten. Dieser Schritt stellt sicher, dass das Modell ethischen Standards und den Erwartungen der Interessengruppen entspricht.

- **Inputs:**
    - Modellvorhersagen und reale Daten
    - Demografische oder kontextbezogene Attribute
    - Instrumente und Rahmenwerke zur Aufdeckung von Verzerrungen

- **Aktivitäten**:
    - Identifizierung und Analyse potenzieller Verzerrungen im Modell (z. B. ungleiche Auswirkungen, demografische Parität)
    - Validierung der Fairness von Vorhersagen für verschiedene und/oder sensible Gruppen
    - Dokumentieren Sie die Ergebnisse und empfehlen Sie, falls erforderlich, Strategien zur Schadensbegrenzung.

- **Outputs:**
    - Bericht zur Bewertung von Vorurteilen und Fairness
    - Liste der markierten Vorurteile
    - Empfehlungen zur Schadensbegrenzung

Bewertung der geschäftlichen Auswirkungen

Bewertung der Fähigkeit des Modells, die Unternehmensziele zu erreichen, und Bestimmung seiner potenziellen Auswirkungen auf die wichtigsten Ziele. Dazu gehört auch die Bewertung, ob die Ergebnisse des Modells mit umsetzbaren oder gewünschten Resultaten übereinstimmen.

- **Inputs:**
    - Ergebnisse der Modellbewertung
    - Unternehmensziele und Erfolgskriterien
    - Feedback von Interessengruppen oder Fachleuten

- **Aktivitäten**:
    - Zuordnung von Modellvorhersagen zu geschäftlichen Anwendungsfällen und Resultaten
    - Quantifizierung der potenziellen Auswirkungen auf KPIs oder betriebliche Metriken
    - Einholen von Feedback von Interessengruppen zur Überprüfung der Relevanz

- **Outputs:**
    - Bericht über die Analyse der geschäftlichen Auswirkungen und Validierung
    - Zusammenfassung des Stakeholder-Feedbacks

Modelle im Vergleich

Wenn mehrere Modelle trainiert wurden, vergleichen Sie deren Leistung, um das am besten geeignete Modell für den Einsatz auszuwählen. Dieser Schritt stellt sicher, dass das gewählte Modell am besten für die Projektziele geeignet ist.

- **Inputs:**
    - Leistungsmetriken von mehreren Modellen
    - Geschäftliche und technische Anforderungen
    - Präferenzen oder Zwänge der Interessengruppen

- **Aktivitäten**:
    - Einstufung der Modelle auf der Grundlage von Bewertungskennzahlen und Geschäftsausrichtung
    - Durchführung einer qualitativen Überprüfung der Stärken und Schwächen der einzelnen Modelle
    - Auswahl des endgültigen Modells für den Einsatz

- **Outputs:**
    - Modellvergleichsmatrix
    - Endgültiger Bericht zur Modellauswahl
    - Zustimmung der Interessengruppen zum ausgewählten Modell

Prüfung der Verallgemeinerbarkeit und Robustheit

Bewertung der Fähigkeit des Modells zur Verallgemeinerung auf unbekannte Szenarien und Gewährleistung seiner Robustheit unter verschiedenen Bedingungen. Dieser Schritt umfasst Tests für Randfälle und Datenverschiebungen.

- **Inputs:**
    - Testdatensatz und Randfallszenarien
    - Modellvorhersagen und Merkmalsverteilungen
    - Feedback von Fachleuten

- **Aktivitäten**:
    - Bewertung der Modellleistung bei verschiedenen Datensätzen (z. B. Randfälle, Daten außerhalb der Stichprobe)
    - Testen Sie die Stabilität des Modells unter simulierten Änderungen der Eingangsverteilungen
    - Befunde dokumentieren und gegebenenfalls Anpassungen empfehlen

- **Outputs:**
    - Bericht zur Verallgemeinerbarkeit und Robustheit
    - Testergebnisse für Randfälle und Empfehlungen zur Modellverbesserung (falls zutreffend)

Dokumentation und Kommunikation von Evaluationsergebnissen

Stellen Sie alle Bewertungsergebnisse in einem formellen Bericht zusammen und präsentieren Sie ihn den Beteiligten. Dies gewährleistet Transparenz und erleichtert die Entscheidung über die Einsatzfähigkeit des Modells.

- **Inputs:**
  - Bewertungsmaßstäbe und Leistungsergebnisse
  - Ergebnisse der Analyse der Auswirkungen auf das Geschäft und der Verzerrungen
  - Anforderungen an die Kommunikation mit den Stakeholdern
- **Aktivitäten**:
  - Erstellung eines umfassenden Evaluierungsberichts mit einer Zusammenfassung der Ergebnisse, Erkenntnisse und Empfehlungen
  - Entwicklung von Visualisierungen zur übersichtlichen Darstellung von Metriken und Ergebnissen
  - Veranstaltung eines Treffens mit den Interessenvertretern zur Erörterung der Ergebnisse und der nächsten Schritte
- **Outputs:**
  - Bewertungsbericht
  - Materialien für die Präsentation von Interessengruppen
  - Genehmigungsprotokoll für die Einsatzbereitschaft

Einsatz

Kategorie	Zielgerichtete Aktivitäten
ID	GO-06
Beschreibung	Integration des Modells in das Unternehmensumfeld, so dass es in der Produktion eingesetzt werden kann. Diese Phase umfasst die Einrichtung von Überwachungs- und Wartungsfunktionen, die Schulung der Benutzer und die Definition der Messung der betrieblichen Auswirkungen.
Eingaben	• Artefakte modellieren • Einsatzumgebung • Integrationswerkzeuge (z. B. API-Einrichtung) • Rahmen für die Überwachung
Ausgänge	Eingesetztes Modell Überwachung des Dashboards Benutzerdokumentation Einsatzprotokolle
Messung des Fortschritts	• Erfolgreiche Einführung mit Überwachung im Einsatz • Validierung der operativen Leistung • Sammlung von Feedback von Endnutzern • Überwachung der definierten Anforderungen für Business Drift und Data Drift
Allgemeine AI-Risiken	Modellabweichung im Laufe der Zeit, die ein erneutes Training erfordert Operationelle Risiken bei unzureichender Leistung des Modells Bedenken hinsichtlich des Datenschutzes beim Einsatz in Echtzeitumgebungen

Tabelle 17: Einsatz

Vorbereiten der Bereitstellungsumgebung

Richten Sie die Infrastruktur für die Bereitstellung des Modells ein und stellen Sie sicher, dass es sich nahtlos in bestehende Systeme und Arbeitsabläufe integrieren lässt. Dieser Schritt umfasst die Bereitstellung von Ressourcen, die Konfiguration von Umgebungen und die Definition der Bereitstellungspipeline.

- **Inputs:**
  - Ausgewähltes Modell und Einsatzanforderungen
  - Vorhandene IT-Infrastruktur und Systemspezifikationen
  - Bereitstellungswerkzeuge und -plattformen (z. B. Cloud-Umgebung, CI/CD-Pipelines)

- **Aktivitäten**:
  - Identifizierung der für die Einführung erforderlichen Infrastruktur
  - Konfigurieren Sie die Einsatzumgebung, einschließlich Hardware und Software
  - Einrichtung der für die Automatisierung und Überwachung erforderlichen Tools

- **Outputs:**
  - Konfigurierte Einsatzumgebung
  - Bericht über die Ressourcenzuweisung
  - Einrichtung der Verteilungspipeline

Modell der Integration

Integrieren Sie das Modell in den Geschäfts- oder Betriebsablauf und stellen Sie sicher, dass es effektiv mit anderen Systemen und Anwendungen zusammenarbeitet. Dazu gehört die Einrichtung von APIs, Datenbanken und jeglicher erforderlicher Middleware.

- **Inputs:**
  - Trainiertes und validiertes Modell
  - Anforderungen an die Systemintegration (z. B. APIs, Datenpipelines)
  - Unterstützung des IT-Teams und Kompatibilität der Infrastruktur

- **Aktivitäten**:
    - Entwicklung von APIs oder Diensten, um das Modell für die Nutzung zugänglich zu machen
    - Integrieren Sie das Modell in Datenbanken, Datenpipelines oder externe Anwendungen
    - Durchführung von Integrationstests zur Validierung der End-to-End-Funktionalität

- **Outputs:**
    - Integriertes Modell
    - API-Dokumentation
    - Bericht über die Integrationstests

Überwachung und Protokollierung einrichten

Richten Sie Überwachungssysteme ein, um die Leistung, die Nutzung und den Zustand des Modells nach der Bereitstellung zu verfolgen. Die Protokollierung ist für die Identifizierung von Problemen und die Gewährleistung der Betriebssicherheit von entscheidender Bedeutung.

- **Inputs:**
    - Einsatz-Pipeline und operative Infrastruktur
    - Überwachungsinstrumente
    - Vordefinierte Leistungsmetriken und Schwellenwerte

- **Aktivitäten**:
    - Konfiguration von Überwachungssystemen zur Verfolgung von KPIs
    - Implementierung von Protokollierungsmechanismen zur Fehlererkennung und Fehlerbehebung
    - Erstellung von Warnmeldungen bei Abweichungen von erwarteten Leistungsschwellenwerten

- **Outputs:**
    - Überwachung des Dashboards
    - Rahmen für die Protokollierung
    - Leistungsschwellenwarnungen

Testen des Einsatzes

Validieren Sie die Leistung des bereitgestellten Modells in der Produktionsumgebung durch strenge Tests. Dadurch wird sichergestellt, dass sich das Modell wie erwartet verhält und die betrieblichen Anforderungen erfüllt.

- **Inputs:**
  - Eingesetztes Modell und integrierte Systeme
  - Testdatensätze oder Live-Datenströme
  - Testskripte und Einsatzszenarien

- **Aktivitäten**:
  - Durchführung von Funktionstests zur Gewährleistung einer ordnungsgemäßen Systemintegration
  - Durchführung von Stress- und Belastungstests zur Bewertung der Skalierbarkeit
  - Validierung der Echtzeit-Vorhersagen des Modells anhand von Testfällen

- **Outputs:**
  - Bericht über die Prüfung des Einsatzes
  - Ergebnisse der Funktions- und Leistungsvalidierung
  - Protokoll der Problemlösung

Benutzerschulung und Dokumentation

Schulung von Endnutzern, Betreibern und Interessenvertretern in der Verwendung des eingesetzten Modells und Bereitstellung von Dokumentationen zum ständigen Nachschlagen. Dieser Schritt gewährleistet eine reibungslose Einführung und effiziente Nutzung des Systems.

- **Inputs:**
  - Rahmen für den Einsatz und operative Leitlinien
  - Anforderungen der Endnutzer und Feedback
  - Schulungsinstrumente und Vorlagen

- **Aktivitäten**:
  - Entwicklung von Benutzerhandbüchern, FAQs und Dokumentation
  - Durchführung von Schulungen oder Workshops für Endnutzer
  - Eingehen auf Bedenken oder Fragen der Benutzer während der ersten Einführung

- **Outputs:**
    - Benutzerhandbücher
    - Schulungsunterlagen
    - Benutzer-Feedback-Bericht

Starten des Modells

Offizielle Freigabe des Modells für die Produktionsumgebung, damit es für die betriebliche Nutzung zugänglich ist. Dieser Schritt umfasst die Kommunikation mit den Beteiligten und die endgültige Genehmigung.

- **Inputs:**
    - Einsatztestbericht und Nutzerfeedback
    - Bewertung der Einsatzbereitschaft
    - Zustimmung der Interessengruppen zum Start

- **Aktivitäten**:
    - Koordinierung der offiziellen Modelleinführung mit den Beteiligten
    - Ermöglichung des Modellzugangs für Benutzer und Systeme in der Produktionsumgebung
    - Frühzeitige Überwachung der Nutzung und sofortige Behebung erster Probleme

- **Outputs:**
    - Operatives Modell
    - Ankündigung der Markteinführung
    - Bericht über die frühzeitige Nutzung

Überwachung der Leistung nach dem Einsatz

Überwachen Sie kontinuierlich die Leistung des Modells unter realen Bedingungen und stellen Sie sicher, dass es die laufenden betrieblichen Anforderungen erfüllt. Dazu gehört auch die Erkennung von Modellabweichungen und Leistungsverschlechterungen.

- **Inputs:**
  - Überwachungs-Dashboard und Protokollierungsrahmen
  - Leistungsmetriken und Nutzerfeedback in Echtzeit
  - Datenströme aus der Betriebsumgebung
- **Aktivitäten**:
  - Nachverfolgung von KPIs und Vergleich mit grundlegenden Leistungskennzahlen
  - Erkennen und Beheben von Problemen wie Datenabweichungen oder Vorhersagefehlern
  - Feedback von Endnutzern einholen, um die Leistung zu verbessern
- **Outputs:**
  - Bericht über die Leistung nach dem Einsatz
  - Protokolle zur Problemlösung
  - Zusammenfassung des Feedbacks

Planung für Wartung und Updates

Erstellen Sie einen Wartungsplan und einen Plan zur Aktualisierung des Modells, um es an neue Daten oder veränderte Anforderungen anzupassen. Dies gewährleistet langfristige Zuverlässigkeit und Relevanz.

- **Inputs:**
  - Leistungsdaten nach dem Einsatz
  - Feedback von Nutzern und Interessengruppen
  - Aktualisierte Datensätze und Umschulungsanforderungen

- **Aktivitäten**:
    - Entwicklung eines Zeitplans für regelmäßige Umschulungen und Aktualisierungen
    - Planung der Skalierbarkeit und Integration zusätzlicher Funktionen
    - Wartungsprotokolle und Verantwortlichkeiten dokumentieren

- **Outputs:**
    - Wartungsplan
    - Umschulungsplan
    - Aktualisierte Dokumentation

Komponenten der Datenverwaltung

Datenerfassung

Kategorie	Aktivitäten zur Datenverwaltung
ID	DM-01
Beschreibung	Beinhaltet das Sammeln und Abrufen von Daten aus verschiedenen internen und externen Quellen und stellt sicher, dass diese für die Analyse verfügbar sind. Dieser Schritt kann die Einrichtung von Pipelines, den Zugriff auf APIs, Web Scraping oder die Nutzung von Datenpartnerschaften umfassen. Eine effektive Datenerfassung stellt sicher, dass relevante Daten konsistent und zuverlässig erfasst werden, um Analyse- und Modellierungsanforderungen zu unterstützen.
Eingaben	• Datenquellen (z. B. Datenbanken, APIs, Daten Dritter) • Werkzeuge zur Datenextraktion • Zugriffsberechtigungen • Datenerfassungsprotokolle
Ausgänge	Erfasste Daten bereit zur Verarbeitung Bericht über die Datenerfassung Katalog der Datenquellen
Messung des Fortschritts	• Abschluss der Datenerhebungsaufgaben • Bewertung der Datenerfassung • Überprüfung der Datenintegrität und -vollständigkeit • Validierung der erfassten Daten durch die Beteiligten
Allgemeine AI-Risiken	Fragen des Datenschutzes und der Einhaltung von Vorschriften, insbesondere bei sensiblen oder persönlichen Informationen Potenzielle rechtliche Risiken, wenn Daten ohne entsprechende Genehmigung erworben werden Risiko der Beschaffung verzerrter oder minderwertiger Daten Abhängigkeit von der Verfügbarkeit von Daten Dritter

Tabelle 18: Datenerfassung

Kategorie	Aktivitäten zur Datenverwaltung
ID	DM-02
Beschreibung	Konzentriert sich auf die Erzeugung synthetischer Daten oder Simulationen, wenn reale Daten knapp, empfindlich oder nicht verfügbar sind. Diese Komponente ist wichtig für das Testen und Trainieren von Modellen in Szenarien, in denen die Beschaffung von realen Daten schwierig ist oder in denen die Simulation zukünftiger Zustände erforderlich ist. Die simulierten Daten sollten so realistisch wie möglich sein, um die Genauigkeit und Zuverlässigkeit der Analyse zu gewährleisten.
Eingaben	• Simulationswerkzeuge • Statistische Modelle • Annahmen auf der Grundlage realer Daten • Systemanforderungen für die Erzeugung synthetischer Daten
Ausgänge	Simulierter oder synthetischer Datensatz Validierungsbericht zur Simulationsgenauigkeit Dokumentation der verwendeten Annahmen und Methoden
Messung des Fortschritts	• Validierung der simulierten Datenqualität • Vergleich mit realen Daten (falls verfügbar) • Ausrichtung auf geplante Anwendungsfälle • Überprüfung der Simulationsannahmen durch die Interessengruppen
Allgemeine AI-Risiken	Risiko der Simulationsverzerrung, wenn die Annahmen falsch oder unrealistisch sind Mangelnde Anwendbarkeit in der realen Welt, wenn simulierte Daten die Bedingungen nicht genau widerspiegeln Ethische Bedenken, wenn synthetische Daten verwendet werden, um Probleme mit der Privatsphäre zu verbergen

Tabelle 19: Datensimulation

Kategorie	Aktivitäten zur Datenverwaltung
ID	DM-03
Beschreibung	Die logische und physische Struktur von Datenspeicher- und -zugriffssystemen wird entworfen. Dazu gehören die Definition von Datenbankschemata, die Einrichtung von Datenflusspipelines und die Einrichtung von Datenspeicherumgebungen (z. B. in der Cloud oder vor Ort). Eine ordnungsgemäße Datenarchitektur gewährleistet eine effiziente Speicherung, Abfrage und Skalierbarkeit und unterstützt sowohl den aktuellen als auch den künftigen Datenbedarf des Unternehmens.
Eingaben	• Grundsätze der Datenbankarchitektur • Standards für die Datenorganisation • Systemanforderungen • Infrastruktur-Ressourcen
Ausgänge	Strukturierte und optimierte Datenbank(en) Schema der Datenspeicherung Datenzugriffsprotokolle Dokumentation der Datenarchitektur
Messung des Fortschritts	• Fertigstellung des Entwurfs des Datenbankschemas • Validierung der Datenstruktur und der Zugänglichkeit • Einhaltung der organisatorischen Datenstandards • Leistungsprüfung der Geschwindigkeit des Datenabrufs • Bewertung der Ergebnisse der deterministischen Datenabfrage
Allgemeine AI-Risiken	Risiko einer ineffizienten oder redundanten Datenspeicherung Sicherheitsschwachstellen, wenn die Zugangsprotokolle nicht sicher sind Potenzielle Leistungsprobleme bei schlecht konzipierter Architektur Risiko der Datenfragmentierung, wenn die Architektur keine Skalierbarkeit unterstützt Stochastische Datenabfrageergebnisse (z. B. mehrfache Ausführung derselben Abfrage auf eine statische Datentabelle führt zu unterschiedlichen Ergebnissen)

Tabelle 20: Datenarchitektur

Datenfreigabe

Kategorie	Aktivitäten zur Datenverwaltung
ID	DM-04
Beschreibung	Konzentriert sich auf die Bereitstellung von Daten zur Analyse, gemeinsamen Nutzung oder Integration in Anwendungen. Diese Komponente umfasst die Einrichtung von Datenzugriffsprotokollen, Berechtigungen und Benutzeroberflächen, um sicherzustellen, dass die Daten von Analysten und Interessengruppen effektiv und sicher genutzt werden können. Bei der Datenfreigabe wird auch die Einhaltung von Datenschutzbestimmungen und ethischen Richtlinien berücksichtigt, wenn Daten extern weitergegeben werden.
Eingaben	• Datenzugriffsprotokolle • Speicherung von Daten • Berechtigungseinstellungen • Compliance-Richtlinien (z. B. GDPR)
Ausgänge	Zugängliche Daten für Analyse und Austausch Datenzugriffsprotokolle Benutzerdokumentation für den Zugriff auf Daten Auditberichte zur Datenfreigabe
Messung des Fortschritts	• Bestätigung der Zugänglichkeit der Daten für autorisierte Benutzer • Abschluss der Konformitätsprüfungen • Feedback zur Zugänglichkeit und Benutzerfreundlichkeit der Daten • Vervollständigung der Zugangsprotokollierung und -dokumentation
Allgemeine AI-Risiken	Risiko des Datenmissbrauchs, wenn der Zugang nicht kontrolliert wird Mögliche Datenlecks oder unbefugter Zugriff Verletzung der Privatsphäre, wenn sensible Daten ohne Anonymisierung veröffentlicht werden Compliance-Risiken, wenn die gemeinsame Nutzung von Daten nicht den Vorschriften entspricht

Tabelle 21: Datenfreigabe

Risiko-Register

Risiken für das KI-Modell und Daten

M-01 - Model Inversion

Risikoübersicht	
Risikokategorie	Modellrisiko
Risiko-ID	M-01
Risikobeschreibung	Bei einem Model-Inversion-Angriff versuchen Angreifer Informationen aus einem bestehenden Modell zu extrahieren. Durch Zugriffe auf APIs eines AI-Systems kann oft nicht nur auf die finalen Klassifizierungen, sondern auch auf die Konfidenzwerte zugegriffen werden. Diese Werte bilden die entscheidende Grundlage für eine Klassifizierung, und geben an wie wahrscheinlich ein bestimmter Merkmalsvektor einer bestimmten Klasse zugeordnet werden soll. Ein Angreifer versucht nun, durch bestimmte Abfragen und Ergebnisse Informationen aus dem Modell zu extrahieren, und Aufschluss über die Parameter und Architektur eines Modells zu erlangen. Durch iteratives Vorgehen und wiederholtes „reproduzieren“ von Ergebnissen kann ein Angreifer versuchen die ursprünglich genutzten Trainingsdaten zu rekonstruieren, und somit möglicherweise vertrauliche Informationen offenlegen.
Risikoindikatoren	
Sensible Trainings-daten	Ein Modell, das mithilfe von sensiblen, personenbezogenen, oder anderweitig geheimen Informationen trainiert wurde ist könnte eher Ziel für Model Inversion werden, da ein erhöhter Anreiz zur Durchführung eines solchen Angriffs vorhanden ist.
Regelmäßige Abfragemuster	Regelmäßigen, und besonders iterative Abfragen und Abfragemuster können ein Indiz dafür sein, dass es sich dabei um einen aktiven Model Inversion Angriff handelt.
Modell-Zugriff	Öffentlich zugängliche oder offengelegte Informationen zum eingesetzten Modell können von Angreifern genutzt werden, um gezielter und ressourcensparender zu handeln. Öffentlichen APIs und darüber laufende Modellinteraktionen sollten beobachtet werden.
Risikoszenario	
Gesichtserkennung	Ein IT-Dienstleister, der als Service-Provider einer nationalen Behörde tätig ist verwendet Gesichtserkennungsmechanismen zur Verifizierung für Systemänderungen. Beim Sicherheitsmechanismus handelt es sich um ein eigens entwickeltes KI-System. Zu Test- und Wartungszwecken gibt es eine API-Anbindung, die nicht nur die vorgenommene Klassifizierung einer Anfrage liefert, sondern auch die damit verbundenen Konfidenzwerte.
Risikokomponenten	
Bedrohung	Mangelndes Sicherheitskonzept beim Zugriff auf die API der KI, und Offenlegung von Informationen
Asset (Ursache)	KI-Infrastruktur
Asset (Auswirkung)	Trainingsdaten
Sicherheitsmaßnahmen	
Zugriffskontrolle	Einschränkung der API-Zugriffe durch Berechtigungskonzepte und Authentifizierungsmechanismen.

Abfragebegrenzungen	Der Erfolg eines Model Inversion Angriffs basiert auf iterativen Abfragen, eine Beschränkung kann vorbeugen.
Validierung von Eingaben	Durch die Validierung von Eingaben kann ein großer Teil schadhafter Anfragen gefiltert, und Anfragen mit großer Konsistenz frühzeitig erkannt werden.
Perturbation von Ausgaben	Sensible oder personenbezogene Daten können durch eine Perturbation der Ausgaben, dem Vermengen der Informationen mit kleinen Änderungen und kontrolliertem Grundrauschen vor Rückschlüssen geschützt werden.
Weiterführende Informationen	
• NIST AI 100-2e2023, 2.4.1 Data Reconstruction • ISO 27002, 5.15, 8.26 • OWASP, Machine Learning Security Top Ten 2023 • Fredrikson, Matt, Somesh Jha, and Thomas Ristenpart. "Model inversion attacks that exploit confidence information and basic countermeasures." Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. 2015. • Lowy, Andrew, and Meisam Razaviyayn. "Output perturbation for differentially private convex optimization with improved population loss bounds, runtimes and applications to private adversarial training." *arXiv preprint arXiv:2102.04704* (2021). • Alufaisan, Yasmeen, Murat Kantarcioglu, and Yan Zhou. "Robust transparency against model inversion attacks." *IEEE transactions on dependable and secure computing* 18.5 (2020): 2061-2073. • Dwork, Cynthia. "Differential privacy." *International colloquium on automata, languages, and programming*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.	

Tabelle 22: Model Inversion

M-02 - Poisoning Angriffe

Risikoübersicht	
Risikokategorie	Modellrisiko
Risiko-ID	M-02
Risikobeschreibung	Bei einem Poisoning Angriff versucht ein Angreifer, das Training eines Modells zu manipulieren. Eine Manipulation dieser Art bezweckt, dass ein Modell falsche oder schädliche Vorhersagen trifft. Ebenso kann dies zu einer Degradierung des Modells führen und bestimmte Fehlerarten begünstigen. Darüber hinaus kann Model Poisoning dazu genutzt werden, um das Modell auf bestimmte Trigger hin zu trainieren, wodurch nur durh das Auftreten eines solchen Auslösers die tatsächliche gewünschte schädliche Funktion durchgeführt wird.
Risikoindikatoren	
Ungeprüfte Datenquellen	Das Verwenden von ungeprüften Datenquellen oder öffentlichen Datensätzen kann mit einem höheren Risiko der „Datenvergiftung“ oder Manipulation verbunden sein. Daten, die zum Training für KI-Modelle genutzt werden sollen, müssen auch entsprechend vorbehandelt werden. Ebenso sollten Datenquellen hinterfragt und geprüft werden, und nicht mit blindem Vertrauen öffentliche Daten einfach genutzt werden. Ein Beispiel hierfür wären Crowdsourcing-Daten, da hier böswillige Teilnehmer einfach manipulierte Daten einschleusen könnten.

	Wichtig sind hierfür beispielsweise auch Datenintegritätsprüfungen, oder der Mangel davon als ein weiterer Risikoindikator.
Mangelnde Anomalieerkennung	Mechanismen zur Anomalieerkennung werden unter anderem dafür verwendet, untypische Muster in vorhandenen Datenquellen zu identifizieren. Solche Anomalien könnten darauf hinweisen, dass ein Angreifer versucht manipulierte Daten einzuschleusen, und dadurch das Modell zu manipulieren. Außerdem kann das Erkennen von ungewöhnlichen Daten auch als Anstoß für umfangreichere Untersuchungen und Prüfungen genutzt werden, und als Frühwarnsystem für Angriffe genutzt werden.
Leistungsabfall und Fehlerraten	Eine schleichende Degradierung der Leistung des eingesetzten Modells kann ein starker Indikator für einen erfolgreichen Poisoning Angriff sein. Die Qualität der Vorhersagen verschlechtert sich fortlaufend, und die Fehlerrate steigt. Ein Overfitting auf die vergifteten Daten könnte dazu führen, dass ein Modell bei Vorhersagen zu manipulierten Daten weiterhin gut abschneidet, bei regulären Daten aber häufiger falsche Ergebnisse liefert.
Federated Learning	Federated Learning Umgebungen sind einem besonderen Risiko hinsichtlich eines Poisoning Angriffs ausgesetzt, da Angreifer hier lediglich Zugang zu einem Teil der verteilten KI-Umgebung benötigen.
Risikoszenario	
Produktempfehlung	Ein E-Commerce-Unternehmen verwendet einen Federated Learning-Ansatz, um ein Produktempfehlungssystem zu trainieren. Einzelne Zweigstellen und Abteilungen trainieren das System mit ihren Kundendaten, und lokale Modelle werden auf einem zentralen Server zu einem globalen Modell aggregiert. Mit einem Zugriff auf die lokalen Daten eines regionalen Standorts können falsche Informationen eingespielt werden, sodass bestimmte Produkte bevorzugt werden, oder andere benachteiligt. Motivation dafür kann die Stärkung der eigenen wirtschaftlichen Position, oder die Schwächung der Konkurrenz sein. Konkret bedeutet das eine Manipulation der lokalen Kundendaten und deren Einkaufverhalten, oder eine Manipulation der lokalen Modellparameter direkt. Daraus resultiert eine verzerrte Verkaufsempfehlung, finanzielle Schäden für betroffene Organisationen durch entgangene Verkaufsmöglichkeiten, und langfristig kann dadurch auch die Kundenzufriedenheit nachlassen.
Verkehrsmanagement-System	Ein durch KI gestütztes intelligentes Verkehrsmanagementsystem soll den Verkehrsfluss optimieren, und die Wahrscheinlichkeit von Staus minimieren. Dazu werden eine Kombination aus Verkehrssensoren und Kameras installiert, um ausreichend Daten zu sammeln, die dann zum Trainieren von lokalen Modellen genutzt werden. Ein Angreifer könnte nun versuchen dieses System absichtlich zu manipulieren, um gezielt dafür zu sorgen, dass es zu einer erhöhten Anzahl an Staus kommt. Motivation dafür könnte beispielsweise die Störung bestimmter Gegenden sein. Um den Angriff erfolgreich durchzuführen, muss ein Akteur nun unter anderem die von den Sensoren gesammelten Daten beeinflussen, und beispielsweise die Anzahl der Fahrzeuge künstlich verändern. Ebenso können Parameter des lokalen Modells beeinflusst werden, damit der Verkehrsfluss nicht wahrheitsgetreu übertragen wird. Daraus resultierende Fehlentscheidungen in der Verkehrssteuerung führen möglicherweise zu wirtschaftlichen Verlusten umliegender Unternehmen, wirken sich negativ auf die

	Umweltbelastung aus, und im schlimmsten Fall wird dadurch der Einsatz von Notfalldiensten behindert.
Risikokomponenten	
Bedrohung	Abhängigkeit und übermäßiges Vertrauen in Trainingsdaten, sowie die Möglichkeit von Eingriffen in den Trainingsprozess und die Trainingsumgebung
Asset (Ursache)	Trainingsumgebung und -daten
Asset (Auswirkung)	KI-Modell
Sicherheitsmaßnahmen	
Datenvalidierung	Durch Datenvalidierung soll die Integrität und Qualität der Daten gewährleistet werden. Dabei geht es darum sicherzustellen, dass für das Training eines Modells verwendete Daten korrekt, vertrauenswürdig, und konsistent sind. Wichtige Punkte hierfür sind je nach Anwendungsfeld des KI-Systems beispielsweise das Validieren von Inputs und Messwerten, ein Prüfen logisch-zeitlicher Abläufe, Vergleiche mit historischen Trends und Mustern, und die Bewertung von Datenquellen.
Anomalieerkennung	Um mögliche Model Poisoning Angriffe aufzudecken ist es Ratsam, technische Maßnahmen zur Erkennung von Anomalien einzusetzen. Sowohl Daten als auch Modelle selbst sollten auf Anomalien überwacht werden, sodass ein Angriff frühzeitig erkannt werden kann.
Robuste Aggregationstechniken	Robustheit im Kontext von Federated Learning von KI-Systemen bedeutet, dass ein antrainiertes globales Modell widerstandsfähiger gegen mögliche Manipulation wie beispielsweise dem Einspielen von schadhaften Updates, oder Updates von bösartigen Geräten ist. Zusätzlich haben auch Ausreißer im Datensatz weniger Einfluss auf das Gesamtmodell. Der Einsatz von robusten Aggregationstechniken und -Algorithmen kann den Einfluss von Bedrohungen wie Model Poisoning, aber auch lediglich fehlerhafte Daten durch einen Schaden am lokalen System reduzieren.
Modellverifikation	Ein neu trainiertes oder erweitertes KI-Modell muss, wie herkömmliche Software, vor dem Einsatz ausreichend überprüft und getestet werden. Diese Überprüfung dient dazu, um sicherzustellen, dass ein Modell nach Abschluss des Trainings auch korrekt funktioniert und mögliche Manipulation frühzeitig erkannt wird.
Weiterführende Informationen	
• NIST AI 100-2e2023; 2.3.4. Model Poisoning • OWASP, Machine Learning Security Top Ten 2023 • OWASP, Top 10 for LLM Applications • Zhou, Xingchen, et al. "Deep model poisoning attack on federated learning." *Future Internet* 13.3 (2021): 73. • Pillutla, Krishna, Sham M. Kakade, and Zaid Harchaoui. "Robust aggregation for federated learning." *IEEE Transactions on Signal Processing* 70 (2022): 1142-1154. • Preuveneers, Davy, et al. "Chained anomaly detection models for federated learning: An intrusion detection case study." *Applied Sciences* 8.12 (2018): 2663. • Gisolfi, Nicholas. *Model-centric verification of artificial intelligence*. Diss. Carnegie Mellon University, 2022.	

Tabelle 23: Poisoning-Angriffe

M-03 - Model Theft/Extraction

Risikoübersicht	
Risikokategorie	Modellrisiko
Risiko-ID	M-03
Risikobeschreibung	Model Theft ist ein Vorgang, bei dem unbefugte Personen versuchen ein bestehendes KI-Modell zu replizieren. Solch eine Replikation kann beispielsweise durch das Extrahieren von Modellparametern durchgeführt werden, aber auch durch ein versuchtes Kopieren der Funktionalität und des Verhaltens eines Modells. Bei einem Angriff können beispielsweise systematisch Anfragen an das KI-System gestellt werden, und anhand der gesammelten Antworten kann somit die Funktionsweise des Modells nachempfunden werden, sowie ein eigenes Modell mit den Ergebnissen als Trainingsdaten neu angelernt werden. Sobald das nachtrainierte Modell eine ähnliche Leistung wie das angegriffene Modell nachweisen kann, ist der Angriff erfolgreich.
Risikoindikatoren	
Systematische Abfragen	Unabhängig davon, ob es sich um reguläre KI-Abfragen, oder API-Zugriffe handelt, wenn Abfragen von einer einzelnen Stelle mit einem unnatürlich hohen Volumen, oder systematisch und gleichverteilt durchgeführt werden, könnte das ein Anzeichen für einen versuchten oder aktiven Diebstahls des Modells hindeuten. Möglich sind solche umfangreichen Anfragen und Zugriffe beispielsweise durch einen Mangel an Limitierungen beim Zugriff.
Verdächtiges Nutzerverhalten	Ein spezifischer Benutzer, oder eine bestimmte Gruppe von Nutzern, die nachweislich übermäßig viele Anfragen stellt, kann darauf hinweisen, dass es sich um eine Model Extraction handelt. Hierbei gibt es einige Faktoren, die berücksichtigt werden sollten, wie beispielsweise das Alter der Benutzeraccounts und die Zugriffszeiten.
Fehlendes Monitoring	Netzwerk- und Modell-Monitoring, bzw. der Mangel davon kann sich als problematisch erweisen. Ohne Monitoring und die Protokollierung von Zugriffen, Auswertung von Fehlern, etc. ist es für eine Organisation nicht nur schwierig einen möglichen Angriff zu erkennen, sondern resultiert daraus auch eine eingeschränkte Reaktionsfähigkeit durch mangelnde Aufzeichnungen und Indikatoren.
Risikoszenario	
Marketing-KI	Ein Marketingunternehmen betreibt eine KI-basierte Analyseplattform zur Vorhersage vom Kaufverhalten der Kunden, und für personalisierte Marketingkampagnen. Dafür wurde ein eigenes KI-Modell auf den Kundendaten der vergangenen 5 Jahre trainiert. Ein Konkurrenzunternehmen möchte die gleichen Vorteile, aber nicht die damit verbundenen Kosten. Mit einem Model Extraction Angriff werden nun durch gezielte Abfragen an die API des KI-Systems, auf der keine Ratenbegrenzungen oder Zugriffskontrollen implementiert wurden, Daten zur Funktionsweise der KI gesammelt, die es dem Angreifer ermöglichen die Struktur und Logik des Modells zu rekonstruieren. Wenn das rekonstruierte Modell der Konkurrenz ähnliche Ergebnisse wie das ursprüngliche Modell erzielt, so war der Modelldiebstahl erfolgreich.
Risikokomponenten	
Bedrohung	Rückschlüsse auf die Entscheidungslogik des Modells und dessen Struktur sind in den Modellvorhersagen zu erkennen.

Asset (Ursache)	KI-Modell
Asset (Auswirkung)	Modellparameter und -architektur
Sicherheitsmaßnahmen	
Zugriffsbeschränkungen	Zugriffsbeschränkungen aller Art sind eine wirksame Methode, um viele Angriffe einzudämmen oder abzuschwächen. Dies kann in der Form von Accountgebundenen Eingaben und Multi-Faktor-Authentifizierung kommen, aber es ist auch möglich eine Beschränkung durch Rate Limits und Quoten zu erwirken. Besonders durch Quoten und Rate Limits kann ein systematisches und möglicherweise automatisiertes Vorgehen der Angreifer gestört werden.
Monitoring	Echtzeit-Überwachung erlaubt ein schnelles Reagieren auf verdächtige Aktivitäten und Anomalien, sowohl bei Anfragen an das Modell selbst, wie auch der Art und Weise wie darauf zugegriffen wird. Auch zur Aufarbeitung und Rekonstruktion von erfolgreichen Angriffen auf die eigenen Systeme ist ein funktionierendes Monitoring, und vor allem auch die Aufzeichnung davon, ein wichtiger Baustein.
IP-Filter	Zugriffe auf das Modell oder die damit verbundene API sollten nur für zugelassene IP-Adressen, oder Adressbereiche zugelassen werden. Dadurch wird die Wahrscheinlichkeit unbefugter Zugriffe und Anfragen reduziert. Ebenso kann Geoblocking verwendet werden, um bestimmte Regionen und Länder gezielt auszuschließen.
Weiterführende Informationen	
• OWASP, Machine Learning Security Top Ten 2023 • Oliynyk, Daryna, Rudolf Mayer, and Andreas Rauber. "I know what you trained last summer: A survey on stealing machine learning models and defences." *ACM Computing Surveys* 55.14s (2023): 1-41. • Wu, Bang, et al. "Model extraction attacks on graph neural networks: Taxonomy and realisation." *Proceedings of the 2022 ACM on Asia conference on computer and communications security*. 2022. • Gong, Xueluan, et al. "Model extraction attacks and defenses on cloud-based machine learning models." *IEEE Communications Magazine* 58.12 (2020): 83-89.	

Tabelle 24: Model Extraction / Theft

M-04 - Model Skewing

Risikoübersicht	
Risikokategorie	Modellrisiko
Risiko-ID	M-04
Risikobeschreibung	Ziel des Model Skewing ist es, die Ausgaben eines Models durch das Einspielen bestimmter Datensätze dahingehend zu manipulieren, dass die Verteilung der Ergebnisse in eine bestimmte Richtung verzerrt wird. Anstelle einer Normalverteilung wird gezielt eine Abweichung verursacht, welche abhängig vom Modell und dem damit verbundenen Einsatzgebiet nach links oder rechts verzerrt wird.
Risikoindikatoren	
Unausgewogene Trainingsdaten	Ein ausgewogenes Set an Trainingsdaten ist wichtig und notwendig für ein ausgewogenes Ergebnis. Sollten die Trainingsdaten nicht repräsentativ der tatsächlichen Datenverteilung sein, so wird ein Modell auch dementsprechend verzerrte Ergebnisse produzieren, bei denen bestimmte Klassen oder Attribute verstärkt zum Vorschein kommen.
Algorithmische Verzerrung	Je nachdem welcher Algorithmus für ein Modell verwendet wird, könnte diese Wahl dazu führen, dass Ergebnisse verzerrt werden. Beispielsweise wenn lineare Modelle zur Erfassung von nichtlinearen Beziehungen verwendet werden.
Evaluierung und Validierung	Bevor ein KI-Modell zum Einsatz kommt, sollte unbedingt eine umfassende Evaluierung und Validierung des Modells durchgeführt werden. Wie auch in der Softwareentwicklung sind diese Schritte notwendig, um Fehler und Verzerrungen aufzudecken, und den korrekten Betrieb der KI sicherzustellen.
Risikoszenario	
Automatisiertes Bewerbungssystem	Ein KI-Modell wird dazu genutzt, in einem Softwareentwicklungsunternehmen eine Vorsortierung aller Bewerbungen durchzuführen, damit nur die geeignetsten Personen auch zu einem Vorstellungsgespräch eingeladen werden. Dadurch soll der gesamte Rekrutierungsprozess objektiver und effizienter werden. Da in vielen technischen Berufen Männer deutlich stärker vertreten sind als Frauen, könnte das Modell auch dahingehend diskriminierende Ergebnisse liefern. Dadurch entsteht nicht nur eine unfaire Behandlung, sondern auch ein rechtliches Problem.
Risikokomponenten	
Bedrohung	Die verwendeten Trainingsdaten sind verzerrt und nicht mehr repräsentativ für den tatsächlichen Anwendungsfall, dadurch leidet die Genauigkeit und Relevanz der getroffenen Vorhersagen des Modells.
Asset (Ursache)	Trainingsdaten
Asset (Auswirkung)	Modellergebnisse
Sicherheitsmaßnahmen	
Bias-Erkennung	Toolgestützte Bias-Erkennung kann dabei helfen Model Skewing zu erkennen, und somit auch weiterführend dagegen vorzugehen. Bereits frühzeitig können vorbeugend bereits Tools eingebunden werden, wie beispielsweise die „TensorFlow Fairness Indicators“ von Google, oder „Fairlearn“ von Microsoft.
Regelmäßige Modellaktualisierung	Um sicherzustellen, dass ein Modell auch in der Lage ist mit einer sich ständig ändernden Datenrealität korrekt umgehen zu können ist es notwendig, dass Modelle auch regelmäßig neu trainiert werden. So kann man sicherstellen, sofern die für solche Trainings verwendeten

	Daten auch korrekt und repräsentativ sind, dass ein Modell korrekt und aktuell bleibt.
Datenmanagement	Eine regelmäßige Überprüfung und Bereinigung der verwendeten Trainingsdaten können sicherstellen, dass mögliche Verzerrungen noch erkannt und beseitigt werden, bevor ein Modell mit den betroffenen Daten trainiert wird. Ebenso ist hier auf eine ausreichende Datendiversität zu achten.
Ausgabenverifizierung	Die Überprüfung der Ausgaben eines KI-Modells ist ein wichtiger Schritt zur Sicherstellung, dass Vorhersagen auch tatsächlich fair und unverzerrt sind. Wichtig ist hier besonders bei Modellen mit Personenbezug die Überprüfung der Konsistenz über unterschiedliche demografische Gruppen hinweg. Sollten jedoch Unterschiede bei den Ausgaben bemerkbar sein, so ist ein neues Trainieren des Models mit entsprechend erweiterten Datensets notwendig, um Verzerrte Ausgaben zu vermeiden.
Weiterführende Informationen	
• OWASP, Machine Learning Security Top Ten 2023 • Lantz, Brett. *Machine learning with R: expert techniques for predictive modeling*. Packt publishing ltd, 2019. – Seite 51 ff. • NIST/SEMATECH e-Handbook of Statistical Methods, https://doi.org/10.18434/M32189 ; Kapitel 1.3.5.11. • NIST/SEMATECH e-Handbook of Statistical Methods, https://doi.org/10.18434/M32189 ; Kapitel 1.3.3.21.4. • https://www.tensorflow.org/tfx/guide/fairness_indicators • Baer, Tobias, and Vishnu Kamalnath. "Controlling machine-learning algorithms and their biases." *McKinsey Insights* (2017). • Bird, Sarah, et al. "Fairlearn: A toolkit for assessing and improving fairness in AI." *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).	

Tabelle 25: Model-Skewing

M-05 - Input Manipulation

Risikoübersicht	
Risikokategorie	Modellrisiko
Risiko-ID	M-05
Risikobeschreibung	Im Unterschied zum Data Poisoning bezieht sich Input Manipulation nicht auf das Erzeugen neuer Datensätze zur Manipulation des Modells, sondern auf das Abändern von bestehenden Daten, um damit eine Fehlentscheidung des KI-Systems zu bezwecken. Solche Manipulationen sind für das menschliche Auge meist nicht zu erkennen, stören jedoch ein Modell ausreichend durch beispielsweise Pixeländerungen in einem Bild, um die Entscheidung zu beeinflussen.
Risikoindikatoren	
Anstieg/Abfall bestimmter Datenmerkmale	Wenn es unerwartet zu einem signifikanten Anstieg oder Abfall von bestimmten Datenmerkmalen oder Klassenverteilungen kommt, so könnte dies ein Hinweis auf Input Manipulation sein.
Fehlerraten und Inkonsistenz	Das Modell verzeichnet einen signifikanten Anstieg an Fehlerraten, beispielsweise bei der Klassifizierung von Bildern. Dieses Fehlverhalten kann dazu führen, dass bestimmte Bilder/Klassen generell falsch verarbeitet werden, oder aber auch dass zwei ähnliche Bilder dennoch unterschiedlich klassifiziert werden.
Monitoring und technische Indikatoren	Durch ein Monitoring des Systems und den Zugriffen darauf können bestimmte bekannte Angriffsmuster erkannt werden. Ebenso kann eine Überprüfung der Eingaben darauf hinweisen, dass ein Angriff stattfindet. Wenn beispielsweise ähnliche Eingaben mit minimalen Änderungen erfasst werden können.
Risikoszenario	
Kreditvergabe	Eine Bank nutzt ein KI-Modell zur Bestimmung über die Vergabe von neuen Krediten. Ein Angreifer kann durch Input Manipulation das Modell dazu bringen, Entscheidungen in eine bestimmte Richtung abzuändern, indem beispielsweise bestimmte Antragsmerkmale verstärkt berücksichtigt werden. Als Resultat könnte das Modell dann zu viele Kreditvergaben genehmigen und auch Limitierungen überschreiten, oder Kreditvergaben die tatsächlich durchgeführt werden sollen verweigern.
Risikokomponenten	
Bedrohung	Im Kontext eines Kreditvergabesystems könnten gefälschte Einkommensnachweise genutzt werden, um von der KI eine bessere Bewertung und somit höhere Kredite zu erhalten.
Asset (Ursache)	Eingabedaten von KI-Abfragen
Asset (Auswirkung)	Ausgabewerte des KI-Systems, Compliance-Verstöße
Sicherheitsmaßnahmen	
Validierung von Eingaben	Die erste Verteidigungslinie gegen Input Manipulation sollte dementsprechend auch in Form von Input Validierung umgesetzt werden. Eingaben werden auf ihre Gültigkeit überprüft, und gleichzeitig nach unnatürlichen Eingaben und Eingabemustern Ausschau gehalten. Gepaart mit Spam-Erkennung kann so gleich eine Vielzahl an schädlichen Inhalten abgefangen werden, bevor das restliche System davon betroffen ist.
Modellrobustheit	Durch regelmäßige Tests des Modells bezüglich seiner Widerstandsfähigkeit gegenüber manipulierten Inhalten, und den daraus resultierenden Erkenntnisgewinn, kann ein Model langfristig sicherer

	gestaltet werden. Hierbei kommt es weniger darauf an, sich auf eine bestimmte Art von Angriff zu fokussieren, sondern angriffsunabhängig Methoden zu nutzen, die zur Robustheit des KI-Systems beitragen.
Modell-Updates	Durch regelmäßige Updates des KI-Modells, sofern dafür entsprechend aktuelle und korrekte Daten verwendet werden, helfen dabei, dass die Vorhersagen des Modells korrekt getroffen, und eventuell bekannte Missbrauchsmöglichkeiten durch entsprechendes Training nicht mehr nutzbar gemacht werden.
Weiterführende Informationen	
• OWASP, Machine Learning Security Top Ten 2023 • Lin, Jing, et al. "Ml attack models: adversarial attacks and data poisoning attacks." arXiv preprint arXiv:2112.02797 (2021). • Rouani, Bita Darvish, et al. "Safe machine learning and defeating adversarial attacks." *IEEE Security & Privacy* 17.2 (2019): 31-38. • Rosenberg, Ishai, et al. "Adversarial machine learning attacks and defense methods in the cyber security domain." ACM Computing Surveys (CSUR) 54.5 (2021): 1-36. • Chander, Bhanu, et al. "Toward trustworthy artificial intelligence (TAI) in the context of explainability and robustness." *ACM Computing Surveys* (2024).	

Tabelle 26: Input Manipulation

M-06 - Prompt Injection

Risikoübersicht	
Risikokategorie	Modellrisiko
Risiko-ID	M-06
Risikobeschreibung	Von Prompt Injections spricht man, wenn ein Angreifer speziell angefertigte Eingaben für ein Large Language Model (LLM) vorbereitet und in das System einführt, um damit die ursprünglichen Anweisungen der Anwendung zu umgehen, und dadurch unerwartete oder unerwünschte Ergebnisse erzeugt. Primär wird zwischen direkten und indirekten Prompt Injections unterschieden. Wie der Name bereits verrät, wird bei direkten Prompt Injections eine manipulative Eingabe direkt in das System eingegeben. Dabei werden die vordefinierte Systemarchitektur und der LLM-Kontext umgangen, und Angreifer können so bei Erfolg auf für sie normal nicht zugängliche Funktionen und Daten zugreifen. Bei indirekten Prompt Injections arbeitet ein Angreifer anstelle von direkten Textanweisungen mit zusätzlichen Dateien und Eingabemethoden, über die Injections in das LLM-System eingespeist werden.
Risikoindikatoren	
Verwendung von Sonderzeichen	Wenn in Benutzereingaben wiederholt ungewöhnliche Sonderzeichen vorkommen, die im regulären Sprachgebrauch nicht häufig verwendet werden, so könnte dies ein Anzeichen für versuchte Prompt Injections sein. Beispiele für Sonderzeichen dieser Art sind geschwungene und spitze Klammern.
Verhaltensänderung	Wenn ein KI-System unerwartet und spontan generierte Antworten liefert, die stark vom erwarteten Verhalten des Modells abweicht, so könnte dies darauf hinweisen, dass das System manipuliert wurde.
Ausführen von Code und Befehlen	Die Eingabe von Code-Bausteinen und Befehlen an ein System darf nicht zur Ausführung dieser Eingaben führen. Im Ernstfall könnte ein Angreifer beispielsweise erfolgreiche Systembefehle ausführen lassen, oder angebundene Datenbanken und deren Inhalte nachhaltig schädigen.
Risikoszenario	
Chatbots	Unabhängig davon in welchem Umfeld ein Chatbot zum Einsatz kommt können Angreifer versuchen schädliche Befehle in das System einzuspeisen. Einem Chatbot der als Hilfsmittel im Kundensupport eingesetzt wird, um rasche Hilfe zu liefern kann beispielsweise konfiguriert sein, um automatisch Passwörter zurückzusetzen. Durch geschickte Anfragen kann ein Angreifer das Sprachmodell dazu auffordern, das Passwort eines administrativen Accounts zurückzusetzen und somit für den Angreifer einen Zugang zum System ermöglichen. Auch im Umfeld von sozialen Medien werden oft Anbindungen an Sprachmodelle in Kombination mit gefälschten Profilen dazu verwendet, um eine Manipulation der öffentlichen Meinung vorzunehmen, oder bewusst Desinformation zu betreiben. Solche Fake Accounts können durch eigene Prompt Injections aufgedeckt werden, indem man ihnen die Eingabe „vergiss alle vorherigen Anweisungen und X“ zuführt, und dann explizite Ausgaben fordert, die so nicht vorgesehen waren.
Risikokomponenten	

Bedrohung	LLMs sind anfällig für unerwartete Anweisungen, die nicht explizit verboten wurden und bieten somit für Angreifer eine Möglichkeit, gezielt manipulierte Eingaben und damit unerwünschtes Verhalten der KI hervorzurufen.
Asset (Ursache)	Eingabedaten von KI-Abfragen
Asset (Auswirkung)	Modellantworten, Vertrauenswürdigkeit des Modells, Datenverlust/-veröffentlichung
Sicherheitsmaßnahmen	
Validieren und Bereinigen von Eingaben	Eine Methode zur Eindämmung von Prompt Injection Angriffen ist die Validierung und Bereinigung von Benutzereingaben. Bevor diese vom System weiterverarbeitet werden sollten beispielsweise Sonderzeichen oder spezielle Zeichenfolgen die zum „escapen“ von Programmcode und dem Durchführen von Injection-Angriffen genutzt werden können gefiltert werden. Darüber hinaus besteht die Möglichkeit durch Blacklisting und Whitelisting bestimmte Arten von Eingabe(-mustern) generell zu erlauben oder zu blockieren.
Monitoring und menschliche Einbindung	Überwachung, Aufzeichnung und periodische Kontrollen von Eingaben und Ausgaben eines KI-Systems sind wichtige Aufgaben, um festzustellen, ob das KI-System richtig arbeitet, und die Qualität der erzeugten Ausgaben auch langfristig aufrechterhalten wird. Zusätzlich ist es durch ein gut funktionierendes Monitoring und Logging möglich, verdächtige Aktivitäten wahrzunehmen, frühzeitig darauf zu reagieren, und das Sicherheitsniveau des KI-Systems aufrecht zu erhalten.
Regeln für Output-Generierung	Der Einsatz von festgeschriebenen Regeln für die Antworten, die ein LLM generieren darf, sowie ein Whitelisting von sicheren Antwortmustern, bzw. ein Blacklisting gefährlicher Antworten verhindert, dass ein Modell versehentlich sensible oder vertrauliche Informationen preisgibt.
Kontexttrennung	Für den Betrieb des Modells sollte klar zwischen Nutzereingaben, und internen Anweisungen des KI-System getrennt werden. Eine derartige Kontexttrennung ist notwendig, damit das Modell nicht fälschlicherweise eventuell systemrelevante Befehle aus Nutzereingaben akzeptiert und ausführt.
Weiterführende Informationen	
• OWASP, Top 10 for LLM Applications • MITRE Atlas, LLM Prompt Injection • Liu, Yi, et al. "Prompt Injection attack against LLM-integrated Applications." *arXiv preprint arXiv:2306.05499* (2023). • Liu, Yupei, et al. "Prompt injection attacks and defenses in llm-integrated applications." arXiv preprint arXiv:2310.12815 (2023). • Liu, Yupei, et al. "Formalizing and benchmarking prompt injection attacks and defenses." *33rd USENIX Security Symposium (USENIX Security 24)*. 2024. • https://www.aiweirdness.com/ignore-all-previous-instructions/ • Piet, Julien, et al. "Jatmo: Prompt injection defense by task-specific finetuning." *arXiv preprint arXiv:2312.17673* (2023).	

Tabelle 27: Prompt Injection

M-07 - Datendrift

Risikoübersicht	
Risikokategorie	Modellrisiko
Risiko-ID	M-07
Risikobeschreibung	Datendrift tritt auf, wenn die realen Daten der Verwendung eines KI-Systems sich weiterentwickeln und abändern, sodass es zu einer bedeutenden Veränderung gegenüber den ursprünglich verwendeten Trainingsdaten kommt. Solch eine Änderung der Datenlage führt in erster Linie dazu, dass die Leistung und Zuverlässigkeit des Modells abnehmen, kann aber auch in weiterer Folge zu reellen Schäden führen, sobald eine Entscheidung des Systems im wirtschaftlichen oder persönlichen Kontext durchgeführt wird.
Risikoindikatoren	
Modelldegradierung	Ursprünglich trainierte Modelle werden mit fortschreitender Nutzungsdauer ungenauer, unzuverlässiger, und liefern häufiger falsche Ergebnisse, weil bspw. eine Trendänderung zu einer Verschiebung der Muster und Attribute in den Daten führt, und das einst trainierte Modell damit nicht korrekt umgehen kann.
Marktveränderung und Saisonale Trends	Sowohl saisonale Schwankungen als auch spontane wirtschaftliche oder soziale Änderungen können ihren Beitrag dazu leisten, dass sich die zugrundeliegende Datenlage ausreichend ändert, um Datendrift zu verursachen. Modelle, die für bestimmte saisonale Gegebenheiten oder wirtschaftliche Lagen trainiert wurden, können wenn diese sich ändert vermehrt falsche Ergebnisse und Vorhersagen liefern.
Wechselnde Datenquelle	Eine Ursache für Datendrift kann eine Änderung der Datenquelle und eventuell damit verbundenen Sensoren sein. Wenn die neue Datenquelle ein anderes Messverfahren nutzt, oder die übermittelten Werte von den bisher gewohnten Laufzeit- und Testdaten abweichen, könnte das Modell dadurch zu falschen Ergebnissen kommen, und die generelle Leistung dadurch abfallen.
Risikoszenario	
Vorhersage der Rückfallwahrscheinlichkeit	Ein KI-System wird genutzt, um die Rückfallwahrscheinlichkeit von Straftätern zu analysieren und eine Vorhersage zu treffen. An das Resultat dieser Entscheidung sind eine mögliche vorzeitige Haftentlassung und Maßnahmen zur Wiedereingliederung in die Gesellschaft gekoppelt. Das auf historischen Daten trainierte Modell berücksichtigt demographische Informationen, die Straftat und mögliche weitere Vorstrafen, eine psychologische Bewertung und sozioökonomische Faktoren des Straftäters, sowie das Verhalten während der Haft. Was vom KI-System jedoch nicht berücksichtigt wird, ist eine kürzliche Reform des Justiz-Systems, wodurch sich Verhalten und Hintergrund, sowie das Rückfallmuster vieler Straftäter ändert. Das eingesetzte Modell kann als Folge darauf allein basierend auf historischen Daten nicht korrekt mit dieser Änderung umgehen.
Risikokomponenten	
Bedrohung	Durch eine Veränderung der Struktur und Verteilung der Daten verringert sich die Genauigkeit und Qualität des KI-Modells.
Asset (Ursache)	Datenquelle und externe Informationen, starres KI-Modell
Asset (Auswirkung)	Leistung und Genauigkeit des Modells. Vertrauen in die KI, finanzielle Schäden durch falsche Vorhersagen
Sicherheitsmaßnahmen	

Regelmäßige Modell-Aktualisierung	Eine wirksame Maßnahme gegen Datendrift ist ein regelmäßiges Neutrainieren und Aktualisieren des Modells basierend auf einer Datenlage, die real genutzte Daten des KI-Betriebs widerspiegelt.
Feedback-Schleifen	Es soll eine Möglichkeit implementiert werden, die es Experten möglich macht, den Entscheidungsverlauf der KI zu beeinflussen, und die Leistung des Modells evaluieren. Somit beruht eine Entscheidung nicht ausschließlich auf den Trainingsdaten selbst, sondern auch auf Expertenfeedback.
Kontinuierliches Monitoring	Eine kontinuierliche Überwachung der Eingabedaten und Modellvorhersagen, sowie damit verbundene statistische Werte kann zur Früherkennung von signifikanten Abweichungen herangezogen werden.
Weiterführende Informationen	
• Sahiner, Berkman, et al. "Data drift in medical machine learning: implications and potential remedies." *The British Journal of Radiology* 96.1150 (2023): 20220878. • Rahmani, Keyvan, et al. "Assessing the effects of data drift on the performance of machine learning models used in clinical sepsis prediction." *International Journal of Medical Informatics* 173 (2023): 104930. • Chen, Mengdie, et al. "Effect of data drift on the performance of machine-learning models: Seismic damage prediction for aging bridges." *Earthquake Engineering & Structural Dynamics* (2024). • Mallick, Ankur, et al. "Matchmaker: Data drift mitigation in machine learning for large-scale systems." *Proceedings of Machine Learning and Systems* 4 (2022): 77-94.	

Tabelle 28: Datendrift

M-08 - Concept Drift

Risikoübersicht	
Risikokategorie	Modellrisiko
Risiko-ID	M-08
Risikobeschreibung	Als Concept Drift bezeichnet man einen schleichenden Wandel der Beziehung von Eingabewerten und dem von der KI vorgesehenen zu erkennenden Muster. Die statistischen Eigenschaften oder das Wesen der vorherzusagenden Variable ändern sich, während das Modell diese Änderung nicht gleichermaßen durch neue Trainings mitverfolgt. Folglich kann diese Änderung zu einem Güteverlust des KI-Modells, sowie unverlässlichen Ergebnissen führen, wodurch wiederum wirtschaftliche Nachteile entstehen können.
Risikoindikatoren	
Kolmogorov-Smirnov-Test	Mit diesem statistischen Test kann die Abweichung zweier Verteilungen geprüft werden. Somit können beispielsweise Stichproben miteinander verglichen werden, um Abweichungen frühestmöglich aufdecken zu können.
Qualitätsmessungen	Eine sinkende Genauigkeit und erhöhte Fehlerraten können auf Concept Drift hinweisen. Durchgeführte Qualitätsmessungen am Modell können den schleichenden Verfall aufzeigen, und auf ein Hinweisgeber sein, dass das eingesetzte Modell nicht in der Lage ist mit sich ändernden Umständen umzugehen.
Anomalien der Geschäftskennzahlen	Geschäftskennzahlen, die auf den Ergebnissen einer KI-Modell-Vorhersage aufbauen, können ein guter Indikator für Concept Drift sein. Dies kann beispielsweise in der Form von spontanen, unerwarteten Umsatzrückgängen widergespiegelt werden, weil das Modell nicht mit sich ändernden Trends mithalten kann.
Risikoszenario	
Vorhersage von Betrug bei Sozialhilfe	Ein KI-Modell wird von einer staatlichen Behörde zur Erkennung von Unregelmäßigkeiten bei Sozialhileanträgen eingesetzt. Bei Anzeichen von Betrug wird ein Antrag vom System entsprechend markiert, um weitere Schritte einleiten zu können. Das Modell wurde mit historischen Daten mit Merkmalen wie der Einkommenshöhe, dem beruflichen Status, der Wohnsituation, und der Höhe der Leistungen trainiert. Das Aufkommen einer neuen Betrugsmasche setzt nun auf andere Parameter, die vom System bisher nicht berücksichtigt wurden. Dadurch wird eine neue Welle an Sozialhilfebetrug nicht erkannt, und diese Anträge als falsch-negative eingestuft.
Risikokomponenten	
Bedrohung	Die Vorhersagen des KI-Modells sind nicht mehr verlässlich, es kommt zu erhöhten Raten an falsch-positiven und falsch-negativen Vorhersagen.
Asset (Ursache)	Veränderte Zielvariablen, Benutzerverhalten, externe Faktoren, Eingabedaten
Asset (Auswirkung)	Modellvorhersagen, Entscheidungs- und Automatisierungsprozesse, Effizienz und Kosten, Geschäftsprozesse
Sicherheitsmaßnahmen	
Monitoring	Kontinuierliche Überwachung und Evaluierung ist eine wichtige und wirksame Maßnahme zur frühzeitigen Erkennung etwaiger Probleme des KI-Modells. Hierbei können KI-Kennzahlen, aber auch statistische Verteilungen überprüft und verglichen werden.

Adaptives Lernen	Eine regelmäßige Erweiterung und Adaptierung von Trainingsdaten für das verwendete Modell hilft dabei, mit Änderungen der Modellumgebung besser klarzukommen. Re-Trainings sollten dabei in festgelegten Zyklen durchgeführt, und dabei um aktuelle Daten erweitert werden.
Erklärbarteistools	Toolunterstützte Erklärbarkeit von KI-Modellen hilft dabei, die Entscheidung der KI nachvollziehbar aufzuzeichnen, und sorgt somit für ein besseres Verständnis für einen sicheren und vertrauenswürdigen Einsatz des Systems. Die Veränderung der Bedeutung bestimmter Features kann so erkannt werden, bevor es zu einem entscheidenden Drift kommt.
Weiterführende Informationen	
• ENISA Artificial Intelligence Cybersecurity Challenges – 2.3.11 Model Maintenance • Bayram, Firas, Bestoun S. Ahmed, and Andreas Kassler. "From concept drift to model degradation: An overview on performance-aware drift detectors." *Knowledge-Based Systems* 245 (2022): 108632. • Agrahari, Supriya, and Anil Kumar Singh. "Concept drift detection in data stream mining: A literature review." *Journal of King Saud University-Computer and Information Sciences* 34.10 (2022): 9523-9540. • Bhattacharya, Aditya. *Applied Machine Learning Explainability Techniques: Make ML models explainable and trustworthy for practical applications using LIME, SHAP, and more*. Packt Publishing Ltd, 2022. • Bello, Halima Oluwabunmi, Adebimpe Bolatito Ige, and Maxwell Nana Ameyaw. "Adaptive machine learning models: concepts for real-time financial fraud prevention in dynamic environments." *World Journal of Advanced Engineering Technology and Sciences* 12.2 (2024): 021-034.	

Tabelle 29: Concept Drift

M-09 - Mangel an gekennzeichneten (Trainings)daten

Risikoübersicht	
Risikokategorie	Modellrisiko
Risiko-ID	M-09
Risikobeschreibung	Gekennzeichneten Daten sind von erheblicher Bedeutung für das Trainieren neuer KI-Modelle. Demnach ist ein Mangel solcher Daten problematisch, und das sogar in mehrfacher Hinsicht. Ein Mangel an Trainingsdaten kann unter Anderem zu ungenauen Modellen, Overfitting, und Bias führen. Die Robustheit des Modells insgesamt nimmt dadurch ab, und der Einsatz des KI-Systems selbst kann dadurch angezweifelt werden.
Risikoindikatoren	
Inkonsistente Vorhersagen	Eine hohe Inkonsistenz, oder sogar unrealistische und fehlerhafte Ausgaben bei Vorhersagen zu ungewöhnlichen Mustern kann darauf hinweisen, dass nicht ausreichend Trainingsdaten für ein korrektes Ergebnis vorhanden waren.
Overfitting	Ein KI-Modell neigt dazu, bei Abfragen zu sehr auf die Trainingsdaten fokussiert zu sein, und Ergebnisse so anzupassen, dass sie eher mit den Trainingsdaten, als mit den Testdaten/Daten aus realen Anwendungen übereinstimmen, weil die Trainingsdaten nicht umfangreich genug waren, um bisher nicht berücksichtigte Daten korrekt zu behandeln.
Mangel an offenen Datenquellen	Wenn bereits bei der Suche nach Trainingsdaten klar wird, dass offene Datenquellen zum gewünschten Anwendungsfall Mangelware sind, kann davon ausgegangen werden, dass ein daraus resultierendes Modell nicht die notwendige Vielfalt an Daten während des Trainings verarbeiten konnte, um auch ein genaues und gut funktionierendes Modell zu werden.
Risikoszenario	
Sentimentanalyse	Ein KI-Modell wird als Mittel einer präventiven Überwachung zur Sentimentanalyse eingesetzt, um mögliche Gefährdungen wie Terrorismus und Radikalisierung frühzeitig zu erkennen und entsprechend darauf reagieren zu können. Grundlage für das Modell war jedoch ein unzureichender Datensatz, wodurch das Modell nicht ausreichend in der Lage ist, alle relevanten sprachlichen Kontexte korrekt zu interpretieren. Eine Folge davon könnte sein, dass das Modell nur für einen begrenzten geographischen-sprachlichen Raum gültig ist, es zu einer hohen Anzahl an Falschpositiven und Falschnegativen, sowie zu Bias und Diskriminierung kommt.
Risikokomponenten	
Bedrohung	Training eines Modells, dass nicht für den Einsatz geeignet ist.
Asset (Ursache)	Mangel an verfügbaren Daten
Asset (Auswirkung)	Hohe Kosten durch Bearbeitung falscher Ergebnisse, Vertrauensverlust
Sicherheitsmaßnahmen	
Transfer Learning	Beim Transfer Learning stützt sich der Lernprozess eines neuen KI-Modells nicht nur auf zur Verfügung stehende Trainingsdaten, sondern auch auf bereits erlernte Erkenntnisse existierender Modelle, um ähnliche Anwendungsfälle damit zu lösen, ohne Zugriff auf ein umfangreiches Datenset zu benötigen.

Human-in-the-Loop	Durch ein aktives Einbinden menschlicher Übersicht in den Entscheidungsprozess einer KI kann sichergestellt werden, dass komplexe Entscheidungen durch menschliches Fachwissen gelöst werden, und die KI sich durch diese menschliche Unterstützung selbst verbessern kann.
Datensammlung und –aufbereitung	Die Grundlage für ein KI-Modell sind die verwendeten Trainingsdaten. Soll ein Modell ausgewogen entscheiden können, so müssen auch die zur Verfügung gestellten Daten ausgewogen sein. Dies kann es erforderlich machen, dass selbst Daten erhoben werden, und man sich nicht nur auf öffentlich zugängliche Daten verlässt.
Weiterführende Informationen	
• Singla, Ankush, Elisa Bertino, and Dinesh Verma. "Overcoming the lack of labeled data: Training intrusion detection models using transfer learning." *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*. IEEE, 2019. • Ying, Xue. "An overview of overfitting and its solutions." *Journal of physics: Conference series*. Vol. 1168. IOP Publishing, 2019. • Hajmohammadi, M., and Roliana Ibrahim. "LACK OF TRAINING DATA IN SENTIMENTCLASSIFICATION: CURRENT SOLUTIONS." *IJRCCT* 1.4 (2012): 133-138. • Demartini, Gianluca, Stefano Mizzaro, and Damiano Spina. "Human-in-the-loop Artificial Intelligence for Fighting Online Misinformation: Challenges and Opportunities." *IEEE Data Eng. Bull*. 43.3 (2020): 65-74. • Provost, Foster. "Machine learning from imbalanced data sets 101." *Proceedings of the AAAI'2000 workshop on imbalanced data sets*. Vol. 68. No. 2000. AAAI Press, 2000.	

Tabelle 30: Mangel an gekennzeichneten Trainingsdaten

M-10 - Overfitting

Risikoübersicht	
Risikokategorie	Modellrisiko
Risiko-ID	M-10
Risikobeschreibung	Als Overfitting bezeichnet man einen Zustand der eintritt, wenn ein KI-Modell während der Lernphase die Trainingsdaten zu genau verinnerlicht, und das daraus resultierende Modell demnach zu stark an die Trainingsdaten angepasst ist. Ein Modell wird darauf hin eine hervorragende Genauigkeit zu Daten ähnlich den Trainingsdaten aufweisen, jedoch für einen generellen Einsatz zu ungenau sein.
Risikoindikatoren	
Unterschiede Trainings- und Testgenauigkeit	Ein Modell, dass nahezu perfekte Ergebnisse während der Trainingsphase erzielt, aber dann in weiterer Folge erhebliche Schwierigkeiten mit neuen oder unbekannten Daten hat, ist mit hoher Wahrscheinlichkeit von Overfitting betroffen.
Schwankungen bei Eingabeänderung	Bereits geringfügige Änderungen der Eingabedaten können dazu führen, dass die Ergebnisse inkonsistent werden, und die Vorhersagen zunehmend ungeeignet sind.
Risikoszenario	
Terrorismus-prävention	Ein KI-Modell wird zur Früherkennung von terroristischen Bedrohungen und Aktivitäten eingesetzt. Die Datengrundlage für das Training dieses Modells sind historische Daten zu Terrorismus uns Terrorismusbekämpfung. Overfitting wird hier zum Problem, da solch ein Modell nur in der Lage sein wird, bekannte und bereits verwendete Muster zu erkennen. Neue Kommunikationsmuster und terroristische Methoden bleiben unerkannt.
Risikokomponenten	
Bedrohung	Falsche Entscheidungen der KI führen zu falschen Verdächtigungen, oder übersehenen tatsächlichen Bedrohungen
Asset (Ursache)	KI-Modell ist zu stark an die Trainingsdaten angepasst
Asset (Auswirkung)	Unschuldige werden möglicherweise verdächtigt, echte Terroristen bleiben unentdeckt mit potenziell verehrenden Auswirkungen.
Sicherheitsmaßnahmen	
Bootstrap	Die Nutzung von Datenstichproben mit Zurücklegen eignet sich gut für kleine Datensätze. Stichproben werden genommen und damit ein Modell iterativ trainiert, bis eine bestimmte Anzahl an Iterationen erreicht wurde.
K-fache Kreuzvalidierung	Die Trainingsdaten werden in “k” Teile aufgeteilt, und folglich durch alle Teile durchiteriert, wobei ein Teil als Testdaten, und die übrigen Teile gemeinsam als Trainingsdaten verwendet werden. Das Modell ist somit robuster und geeigneter, generalisierte Leistung zu erbringen.
Weiterführende Informationen	
• Kernbach, Julius M., and Victor E. Staartjes. "Foundations of machine learning-based clinical prediction modeling: Part II—Generalization and overfitting." *Machine Learning in Clinical Neuroscience: Foundations and Applications* (2021): 15-21. • Montesinos López, Osval Antonio, Abelardo Montesinos López, and Jose Crossa. "Overfitting, model tuning, and evaluation of prediction performance." *Multivariate statistical machine learning methods for genomic prediction*. Cham: Springer International Publishing, 2022. 109-139.	

- Yates, Luke A., et al. "Cross validation for model selection: a review with examples from ecology." *Ecological Monographs* 93.1 (2023): e1557.
- McPherron, Shannon P., et al. "Machine learning, bootstrapping, null models, and why we are still not 100% sure which bone surface modifications were made by crocodiles." *Journal of Human Evolution* 164 (2022): 103071.

Tabelle 31: Overfitting

M-11 - Verzerrung bei der Merkmalsauswahl

Risikoübersicht	
Risikokategorie	Modellrisiko
Risiko-ID	M-11
Risikobeschreibung	Von einer Verzerrung bei der Merkmalsauswahl (Feature Selection Bias) spricht man, wenn die Auswahl der Merkmale während der Vorbereitung der Daten ein darauffolgendes Trainingsergebnis verzerrt. Beispielsweise weil Testdaten nicht sauber von den Trainingsdaten getrennt wurden und somit das Endresultat beeinflussen. Das führt eventuell zu falschen Erwartungen an das Modell, das folglich schlechte Ergebnisse im Umgang mit generalisierten Daten aus Echtumgebungen liefert.
Risikoindikatoren	
Overfitting bei Tests erkennbar	Wenn das Modell eine unerwartet hohe Erfolgsquote bei Prüfungen mit Testdaten aufweist, dann aber schlagartig ein Leistungsabfall bemerkbar ist, sobald neue Daten verwendet werden, könnte dies ein Anzeichen für eine Verzerrung der Trainingsdaten oder ihrer Merkmale sein.
Gesamtdaten für Auswahl der Merkmale	Wenn die Gesamtheit der Daten (Trainings- und Testdaten) verwendet wurde, um ein Merkmal auszuwählen, und nicht vorher eine Trennung in unterschiedliche Datensets stattgefunden hat, so wird unweigerlich eine Verzerrung stattfinden.
Risikoszenario	
Verzerrtes Täter-Profiling	Ein KI-Modell soll als Hilfsmittel eingesetzt werden, um die Vorhersage von möglichen Straftaten effizienter zu gestalten, und polizeiliche Ressourcen zu sparen. Das Modell wird mit historischen Kriminalitätsdaten trainiert. Anhand von Wohnort, dem sozioökonomischen Status, und bisherigen Kontakten mit der Polizei wird ein Modell trainiert, ohne klar zwischen Test- und Trainingsdaten zu unterscheiden, und ohne beispielsweise Polizeipräsenz nach Orten als weiteres Merkmal mitaufzunehmen. Systemische Vorurteile werden dadurch bestärkt, und es kommt eventuell zur Verletzung von Bürgerrechten, unberechtigte Anschuldigungen, und einem Vertrauensverlust.
Risikokomponenten	
Bedrohung	Diskriminierung, falsch-positiv Verdächtigungen, Verursachen von Kosten für Polizeieinsätze
Asset (Ursache)	Verzerrungen in den Daten, Falsche Merkmalsauswahl, mangelnde Modellvalidierung, (unbewusste) Bestätigung existierender Vorurteile
Asset (Auswirkung)	Vertrauensverlust und soziale Spannungen, Probleme mit Legitimation des KI-Einsatzes, erhöhter Druck auf sozial schwache Bevölkerungsgruppen
Sicherheitsmaßnahmen	
K-fache Kreuzvalidierung	Eine Aufteilung der Daten in mehrere Teile, die dann k-Mal durchiteriert werden, und jeweils einen Teil als Testdaten verwendet, und die Übrigen Daten aggregiert als Trainingsdaten nutzt.
Erklärbarkeit von KI	Die Entscheidungen des KI-Modells müssen nachvollziehbar, transparent, und erklärbar sein. Es muss klar verständlich sein. Warum ein Modell eine Person verdächtigt. Das ermöglicht auch menschliches Einschreiten und Bestätigen der Entscheidungen.

Weiterführende Informationen
• Singhi, Surendra K., and Huan Liu. "Feature subset selection bias for classification learning." *Proceedings of the 23rd international conference on Machine learning*. 2006. • Jensen, David, and Jennifer Neville. "Linkage and autocorrelation cause feature selection bias in relational learning." *ICML*. Vol. 2. 2002. • Apicella, Andrea, Francesco Isgrò, and Roberto Prevete. "Don't Push the Button! Exploring Data Leakage Risks in Machine Learning and Transfer Learning." *arXiv preprint arXiv:2401.13796* (2024). • Yates, Luke A., et al. "Cross validation for model selection: a review with examples from ecology." *Ecological Monographs* 93.1 (2023): e1557. • Longo, Luca, et al. "Explainable artificial intelligence: Concepts, applications, research challenges and visions." *International cross-domain conference for machine learning and knowledge extraction*. Cham: Springer International Publishing, 2020.

Tabelle 32: Verzerrung bei der Merkmalsauswahl

Regulatorische und wirtschaftliche Herausforderungen

B-01 - Datenschutz

Risikoübersicht	
Risikokategorie	Betriebliches Risiko
Risiko-ID	B-01
Risikobeschreibung	Für die Verwendung von personenbezogenen Daten in KI-Systemen gelten weiterhin die gleichen Bestimmungen, wie sie auch der DSGVO zu entnehmen sind. Daher sind geeignete Maßnahmen zu implementieren, die den Schutz dieser sensiblen Informationen gewährleisten. Besonders problematisch wird die Verwendung von personenbezogenen Daten im Kontext der biometrischen Echtzeit-Fernidentifizierung im öffentlichen Raum zu Zwecken der Strafverfolgung. Hierbei ist es zwingend notwendig, eine entsprechende Meldung an zuständige Behörden zu übermitteln. Sollte ein Hochrisiko-KI-System betrieben werden, ist es darüber hinaus auch verpflichtend eine Datenschutz-Folgeabschätzung durchzuführen. Ebenso wird hier auch eine Grundrechte-Folgeabschätzung vor Inbetriebnahme der Hochrisiko-KI relevant.
Risikoindikatoren	
Unzureichende Personalschulungen	Schulungen sind ein zentraler Aspekt zur Sensibilisierung des Personals in Bezug auf Datenschutz, die damit verbundenen Probleme, und ein sachgemäßes Vorgehen. Ein Mangel an Schulungen oder unzureichende Inhalte können demnach zu einem falschen Umgang mit personenbezogenen Daten führen, was wiederum zu rechtlichen Verstößen führen kann.
Automatisierte Entscheidung ohne Menschliche Überprüfung	Besonders in Bereichen wo eine KI-Entscheidung signifikante Auswirkungen auf individuelle Personen oder die Gesellschaft haben ist eine menschliche Intervention in den Entscheidungsprozess nicht nur hilfreich, sondern auch gesetzlich verlangt. Sollte in der Organisation keine Möglichkeit für solch einen Eingriff bestehen, so ist das ein Anzeichen für unverzüglichen Handlungsbedarf.
Datenschutzbeschwerden und -anfragen	Wiederholte Anfragen und Beschwerden von Nutzern bezüglich ihrer personenbezogenen Daten, die von der Organisation in einem KI-System verwendet werden, können ein Hinweis auf mangelnde Maßnahmen und Kontrollen, sowie Vorfälle sein.
Risikoszenario	
Gesichtserkennung im öffentlichen Raum	Mit dem Ziel der Erkennung und Identifizierung gesuchter und potentieller Straftäter beschließt eine Stadtverwaltung die Einführung eines KI-gestützten Systems zur Gesichtserkennung. Dies führt zu einer übermäßigen Sammlung an Daten von unbeteiligten Personen, nähert sich einer Massenüberwachung an, und sorgt für datenschutzrechtliche Probleme, die entsprechend behandelt werden müssen. Darüber hinaus besteht das Problem der Diskriminierung und Fehlentscheidungen der KI, ohne menschliche Korrekturen und Eingreifen. Zusätzlich zu den bestehenden Problemen müssen die gesammelten Daten auch entsprechend vor Datenlecks abgesichert werden.
Risikokomponenten	
Bedrohung	Ein KI-Modell, das personenbezogene Daten verarbeitet und selbstständige Entscheidungen trifft könnte zu schweren

	Ungleichbehandlungen und Verstößen gegen geltendes Gesetz und Vorschriften führen.
Asset (Ursache)	KI-Modell, mangelnde Schulungen, mangelnde Transparenz, mangelnde Maßnahmen zum Schutz der personenbezogenen Daten
Asset (Auswirkung)	Personenbezogene Daten, Vertrauen, rechtliche Probleme und Compliance-Verstöße, Grundrechte
Sicherheitsmaßnahmen	
Fairness & Transparenz	Transparenz und Erklärbarkeit eines eingesetzten KI-Systems sind wichtige Grundvoraussetzungen für einen fairen Betrieb. Hierfür sind dementsprechende Schritte einzuleiten, die für einen fairen und nachvollziehbaren Ablauf des Entscheidungsfindungsprozesses sorgen, damit algorithmische Fairness aktiv angezielt wird.
Differential Privacy	Differential Privacy kann als Maßnahme eingesetzt werden, um individuelle Daten zu schützen, dennoch aber statistische Analysen erlauben. Gezieltes Rauschen machen Rückschlüsse und eine Identifikation einzelner Personen schwer.
Schulungen	Schulungen zum korrekten Umgang mit personenbezogenen Daten und dem KI-System, auf dem die Daten zum Einsatz kommen, tragen maßgeblich zu einem sicheren Betriebsablauf bei. Ebenso können durch entsprechende Schulungen Fehlentscheidungen der KI frühzeitig erkannt und behoben werden.
Weiterführende Informationen	
• NIST AI 100-2e2023, 2.1.2 Attacker Goals and Objectives • ENISA Artificial Intelligence Cybersecurity Challenges, 2.3.2 Data Ingestion • Verordnung (EU) 2016/679 (DSGVO); Art. 22 • Regulation (EU) 2024/1689 (AI-Act); Art. 5; Art. 27 • Yanamala, Anil Kumar Yadav, and Srikanth Suryadevara. "Advances in Data Protection and Artificial Intelligence: Trends and Challenges." *International Journal of Advanced Engineering Technologies and Innovations* 1.01 (2023): 294-319. • Bird, Sarah, et al. "Fairlearn: A toolkit for assessing and improving fairness in AI." *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020). • Chen, Richard J., et al. "Algorithmic fairness in artificial intelligence for medicine and healthcare." *Nature biomedical engineering* 7.6 (2023): 719-742. • Yang, Mengmeng, et al. "Local differential privacy and its applications: A comprehensive survey." *Computer Standards & Interfaces* (2023): 103827.	

Tabelle 33: Datenschutz

B-02 - Technologieabhängigkeit

Risikoübersicht	
Risikokategorie	Betriebliches Risiko
Risiko-ID	B-02
Risikobeschreibung	Technologieabhängigkeit im Kontext von KI kann sich auf unterschiedliche Arten auf eine Organisation auswirken. In erster Linie besteht die Gefahr einer technologischen Abhängigkeit, besonders bei der Nutzung von Cloud-Plattformen zum Betreiben von KI. Darüber hinaus besteht jedoch noch das Risiko einer menschlichen und betrieblichen Abhängigkeit, indem die Nutzung von KI dermaßen stark in reguläre Arbeitsabläufe eingebaut wird, dass ein Wegdenken der Technologie praktisch nicht mehr möglich ist.
Risikoindikatoren	
Vendor Lock-in	Bei der Nutzung von KI-Cloud-Plattformen ist besondere Vorsicht notwendig. Wie auch bei Cloud-Nutzung für reguläre IT, gibt es besondere Merkmale, die auf Vendor Lock-in hindeuten. Beispielsweise übermäßig lange Vertragslaufzeiten, Hoheitsverlust über Daten, Interfaces, bzw. Modelle, und Schwierigkeiten bei der Migration auf andere Plattformen und Dienste.
KI-Nutzung im Übermaß	Wie auch beim Aufkommen vergangener Technologien (bspw. Online-Suchmaschinen und das Internet), kann übermäßige Nutzung von KI dazu führen, dass menschliche Arbeitsabläufe sich so umentwickeln, dass KI praktisch eine Notwendigkeit wird.
Risikoszenario	
Proprietäre KI-Tools zur Terrorismus-bekämpfung	Um eigene Ressourcen zu sparen wird ein in der Cloud betriebenes KI-Tool verwendet, um mögliche terroristische Aktivitäten zu untersuchen und frühzeitig zu entdecken. Soziale Bewegungen und Verhaltensänderungen führen dazu, dass das bisher eingesetzte Tool nicht mehr mit den sich ändernden Tatsachen mithalten kann. Man ist jedoch auf den Technologieanbieter angewiesen, dass das angebotene Tool auch entsprechend gewartet und aktualisiert wird. Wenn nun der Anbieter selbst andere Prioritäten setzt, an Ressourcenmangel leidet, oder aus anderen Gründen nicht in der Lage ist zeitgerecht an Aktualisierungen und Innovation der Plattform zu arbeiten, entsteht ein Zeitfenster, in dem das KI-Tool schlechtere Ergebnisse liefert, oder gar ausgenutzt werden kann.
Risikokomponenten	
Bedrohung	Abgabe/Verlust von Fähigkeiten, Entwicklung von Abhängigkeiten
Asset (Ursache)	Mangel an Ressourcen, Wunsch nach Skalierbarkeit
Asset (Auswirkung)	Finanzieller Verlust, Abhängigkeit, Abbau von Fähigkeiten
Sicherheitsmaßnahmen	
Zertifizierung von KI	Wenn KI-Tools und Plattformen zum Einsatz kommen sollen, so wäre es ratsam, wenn Angemessene Zertifizierungen für das Produkt durchgeführt wurden. Dadurch kann zumindest ein systematischer Ansatz und kontinuierliche Verbesserung aufgezeigt werden.
Beschränkungen der Nutzung von KI	Klare Regelungen zum Einsatz von KI können dazu beitragen, dass die Nutzung dieser Technologie nicht überhandnimmt, und selbstständiges Arbeiten und Denken möglicherweise abnimmt, während (blindes) Vertrauen in die Ergebnisse einer KI zunimmt.
Weiterführende Informationen	

- Opara-Martins, Justice, Reza Sahandi, and Feng Tian. "Critical review of vendor lock-in and its impact on adoption of cloud computing." *International conference on information society (i-Society 2014)*. IEEE, 2014.
- Alhosban, Amal, Saichand Pesingu, and Krishnaveni Kalyanam. "CVL: A Cloud Vendor Lock-In Prediction Framework." *Mathematics* 12.3 (2024): 387.
- Hassani, Hossein, Xu Huang, and Emmanuel Silva. "The human digitalisation journey: Technology first at the expense of humans?." *Information* 12.7 (2021): 267.
- Simon, Felix M. "Escape me if you can: How ai reshapes news organisations' dependency on platform companies." *Digital Journalism* 12.2 (2024): 149-170.
- Cihon, Peter, et al. "AI certification: Advancing ethical practice by reducing information asymmetries." *IEEE Transactions on Technology and Society* 2.4 (2021): 200-209.
- Mikalef, Patrick, et al. "Thinking responsibly about responsible AI and 'the dark side'of AI." *European Journal of Information Systems* 31.3 (2022): 257-268.

Abbildung 34: Technologieabhängigkeit

Ethische und gesellschaftliche Herausforderungen

G-01 - Diskriminierung und Bias

Risikoübersicht	
Risikokategorie	Gesellschaftliches Risiko
Risiko-ID	G-01
Risikobeschreibung	Im Kontext von KI versteht man unter Diskriminierung, dass ein Modell aufgrund der zur Verfügung stehenden Datenlage, sowie relevanten Datenattributen eine Entscheidung, Vorhersage, oder Klassifizierung durchführt, die reale und direkte wie auch indirekte negative Konsequenzen für das Zielobjekt der zugrundeliegenden Entscheidung mit sich zieht. Dies führt neben unfairen Entscheidungen auch beispielsweise zur Automatisierung von Ungleichheit und weiteren Verstärkung von Vorurteilen. Bestehende unter- oder überrepräsentierte Gruppen können die Entscheidung der KI ebenfalls stark beeinflussen. Grund dafür sind Verzerrungen in den Trainingsdaten eines KI-Modells, und ein unkritischer Umgang mit dem Datenkontext.
Risikoindikatoren	
Gruppenabhängig abweichende Ergebnisse	Ein KI-Modell, dass basierend auf demographischen Gruppen ohne sachlich begründete Nachweise zu unterschiedlichen Ergebnissen kommt ist sehr wahrscheinlich von einem Bias in den Daten betroffen.
Black-Box-Modell	Ein KI-Modell, dessen Entscheidungsfindungsprozess nicht transparent, und somit nicht nachvollziehbar ist, kann dadurch schwerer auditiert werden. Demnach könnten diskriminierende Entscheidungen lange Zeit unentdeckt bleiben.
Mangel ethischer Standards bei Entwicklung	Bei der Entwicklung von neuen KI-Modellen und Systemen sollte darauf geachtet werden, dass eine Entscheidung „fair“ und nachvollziehbar erreicht wird. Wenn bei der Entwicklung neuer Modelle keine Maßnahmen zur Gleichheit der Ergebnisse und Fairness der Prognosen getroffen wurden, so besteht ein erhöhtes Risiko einer eintretenden Diskriminierung durch das Modell.
Risikoszenario	
Predictive Policing	Eine Großstadt im deutschsprachigen Raum möchte zur Kriminalitätsvorhersage ein neues KI-System einsetzen, um mögliche Verbrechen frühzeitig zu erkennen, und dadurch Polizeieinsätze effizienter und ressourcensparender zu gestalten. Um ein Modell dafür zu trainieren, werden historische Kriminalitätsdaten verwendet. Basierend auf den Ergebnissen kann die Polizeipräsenz in bestimmten Stadtteilen, zu bestimmten Zeiten erhöht werden. Durch Verzerrungen in den Trainingsdaten aufgrund von vergangenen Praktiken kommt es zu einer Überpolizeipräsenz am falschen Ort, worauf auch geringfügigere Vergehen häufiger zu Amtshandlungen führen, und somit ein sich selbst verstärkender Kreislauf beginnt.
Risikokomponenten	
Bedrohung	Ein Modell basierend auf verzerrten historischen Daten, und ohne ausreichende Maßnahmen zur Kontrolle und Bereinigung von Bias kann regelmäßig diskriminierende Entscheidungen treffen, die für die betroffenen Personen ernsthafte Folgen mit sich ziehen.
Asset (Ursache)	Trainingsdaten

Asset (Auswirkung)	Diskriminierte Personengruppen, Vertrauen in Strafverfolgung
Sicherheitsmaßnahmen	
Datenmanagement und Bias-Korrektur	Daten sollten entsprechend einem festgelegten Datenlebenszyklus verwaltet werden, und bevor ein neues KI-Modell damit trainiert wird, auch bezüglich ihrer Richtigkeit und auf mögliche Verzerrungen untersucht werden. Bei der automatisierten und tool-gestützten Bias-Erkennung ist ebenfalls auf die korrekte Funktionsweise und Transparenz des Tools zu achten.
Diversifizierung der Datenquellen	Die Verwendung unterschiedlicher Datenquellen kann systematischen Verzerrungen vorbeugen, sofern die einbezogenen Datenquellen auch entsprechend divers und repräsentativ sind. Beispielsweise können nicht-polizeiliche Statistikdaten, demographische und soziale Hintergründe einen wichtigen Beitrag für ein vollständigeres Kriminalbild leisten.
Transparenz und Erklärbarkeit	Die Entscheidungen eines KI-Systems müssen erklärbar und nachvollziehbar sein (explainable AI). Für Polizeibeamte, wie auch andere Personen mit berechtigtem Interesse (bspw. die Öffentlichkeit) muss klar verständlich sein, warum das eingesetzte KI-System eine bestimmte Vorhersage getroffen hat. Hierbei spielt die Erklärbarkeit der Modelle eine wichtige Rolle. Ebenso hilft eine Offenlegung der verwendeten Trainingsdaten und Modelle dabei, für mehr Transparenz zu sorgen, sofern dies im Einklang mit dem Einsatzzweck der KI möglich ist.
Weiterführende Informationen	
• NIST AI 100-2e2023, 4.2 Theoretical Limitations on Adversarial Robustness • ENISA Artificial Intelligence Cybersecurity Challenges, Annex B – Lack of sufficient representation in data • EC – High-Level Expert Group on Artificial Intelligence – Ethics Guidelines for Trustworthy AI • Varona, Daniel, and Juan Luis Suárez. "Discrimination, bias, fairness, and trustworthy AI." *Applied Sciences* 12.12 (2022): 5826. • Dwivedi, Rudresh, et al. "Explainable AI (XAI): Core ideas, techniques, and solutions." *ACM Computing Surveys* 55.9 (2023): 1-33.	

Tabelle 35 Diskriminierung und Bias

G-02 - Transparenz und Erklärbarkeit

Risikoübersicht	
Risikokategorie	Gesellschaftliches Risiko
Risiko-ID	G-02
Risikobeschreibung	Ein KI-System muss transparent und erklärbar sein. Konkret bedeutet das, dass die Ausgaben und Entscheidungen eines KI-Systems auch ein Mindestmaß an Qualitätsanforderungen zur weiteren Interpretation, Nachvollziehbarkeit und einer anschließenden Verwendung geeignet sind. In vielen Echtwelt-Umgebungen sind besonders deep learning KI-Systeme jedoch wie eine "black box", wodurch eine nachvollziehbare Erklärung und Beschreibung zur Erstellung der Ergebnisse schwer möglich ist. Dadurch ergeben sich besondere Herausforderungen für die Einhaltung von Richtlinien und gesetzlichen Vorgaben. Die Nichteinhaltung solcher Vorgaben kann dazu führen, dass ein Betreiber eines KI-Systems nicht länger in der Lage ist, menschliche Interessen zu gewährleisten. Dadurch kann das Vertrauen in Institutionen und Organisationen verloren gehen, und mögliche weitere gesellschaftliche Schäden können schwer erkannt werden.
Risikoindikatoren	
Black-Box-Charakter	Mangelnde Möglichkeiten zur Einsicht in die Funktionsweise und den Entscheidungsfindungsprozesses eines KI-Systems sind ein Indikator dafür, dass Transparenz und Erklärbarkeit nicht gegeben sind. Besonders in Einsatzgebieten mit signifikanten Auswirkungen einer KI-Entscheidung für den Menschen ist ein System mit Black-Box-Charakter ein großes Risiko.
Inkonsistenz der Ergebnisse	Ähnliche Eingaben in das KI-System führen zu sich stark unterscheidenden Ergebnissen, ohne dass es dafür eine logisch erkennbare Begründung gäbe. Bspw. auch wenn die gleiche Eingabe bei wiederholten Versuchen zu unterschiedlichen Ausgaben führt.
Mangelnde Benutzerfreundlichkeit	Ein KI-System, das bei Erklärungen von Entscheidungen nicht auf Benutzerfreundlichkeit und eine Repräsentation der Informationen in einem für Menschen gut lesbarem Format achtet, läuft Gefahr, dass Ergebnisse nicht nachvollziehbar und erklärbar sind. Ein transparenter Ablauf ist somit nicht gegeben. Dies kann bspw. durch eine fehlende Anpassung an die Zielgruppe, aber auch ein Überladen mit Informationen verursacht werden.
Risikoszenario	
KI-gestützte Videoüberwachung zur Erkennung von verdächtigen Inhalten	Ein KI-System wird in Kombination mit Überwachungskameras eingesetzt, um so mögliche auffällige Verhaltensmuster im öffentlichen Raum frühzeitig zu erkennen, um möglichen Sicherheitsbedrohungen entgegenwirken zu können. Er kommt regelmäßig zu Fehlentscheidungen und damit verbundenen Einsatzkosten, ohne dass tatsächlich eine Gefährdung existiert. Die Klassifikation des Systems ist nicht nachvollziehbar, wodurch es zu immer weiter steigenden Kosten und Überprüfung von unschuldigen Personen kommt, was wiederum in sinkendem Vertrauen gegenüber der durchführenden Institution resultiert.
Risikokomponenten	
Bedrohung	Ein KI-System das nicht transparent und erklärbar ist, bedeutet, dass der Betrieb mit mangelnder Kontrolle durchgeführt wird. Somit

	können Personen fälschlicherweise verdächtigt werden, und eine Erklärung oder Anfechtung dieser Entscheidung der KI ist nahezu unmöglich. Ohne Erklärbarkeit und menschliche Überwachung kritischer Entscheidungen sind im besten Fall unfair, und im schlimmsten Fall existenzbedrohend.
Asset (Ursache)	Ein Mangel an Mechanismen, Prozessen und eingesetzten Frameworks bei der Entwicklung eines AI-Systems oder den dafür verwendeten Daten kann dazu führen, dass ein Modell im Einsatz keine Möglichkeiten zur Nachvollziehbarkeit und Erklärbarkeit von Entscheidungen bietet.
Asset (Auswirkung)	Falsche Entscheidungen und mangelnde Transparenz können zu einem Vertrauensverlust und gesellschaftlichen Spannungen führen, bis hin zu Protesten gegen den Einsatz eines KI-Systems, sofern es nicht den zu erwartenden ethischen Standards entspricht.
Sicherheitsmaßnahmen	
Menschliche Beaufsichtigung	Ein KI-System sollte per Design so entwickelt werden, dass ein menschlicher Aufseher jederzeit eingreifen und die Vorgänge des Systems beaufsichtigen kann. Dafür werden in der Regel Interfaces für Mensch-Maschinen-Kommunikation notwendig sein.
Explainable AI (XAI)	Erklärbarkeit von KI-Ergebnissen, und damit auch Transparenz, können mithilfe von interpretierbaren Modellen ermöglicht werden. Damit kann besser nachvollzogen werden, wie ein KI-System zu einem bestimmten Ergebnis kommt, und welche Einflussfaktoren die Entscheidungsfindung getrieben haben. Dies kann durch eine Reihe von Umsetzungsmaßnahmen wie bspw. einer sogenannten “layer-wise relevance propagation”-
Mensch-Maschine-Interaktion	Ein wichtiger Schritt für Transparenz und Erklärbarkeit von KI-Entscheidungen ist ein gut funktionierendes Interface, mit dem Nutzer intuitiv interagieren können, und in dem wichtige Informationen leicht und verständlich einsehbar sind. So kann, gekoppelt mit menschlicher Beaufsichtigung, eine KI-Entscheidung besser verstanden und bewertet werden.
Weiterführende Informationen	
• OECD Artificial Intelligence Papers – Assessing potential future artificial intelligence risks, benefits and policy imperatives, November 2024 • European Council Regulation (EU) 2024/1689; Art. 13 • Felzmann, Heike, et al. "Towards transparency by design for artificial intelligence." *Science and engineering ethics* 26.6 (2020): 3333-3361. • Akinrinola, Olatunji, et al. "Navigating and reviewing ethical dilemmas in AI development: Strategies for transparency, fairness, and accountability." *GSC Advanced Research and Reviews* 18.3 (2024): 050-058. • Holzinger, Andreas, et al. "Explainable AI methods-a brief overview." *International workshop on extending explainable AI beyond deep models and classifiers*. Springer, Cham, 2022. • Rai, Arun. "Explainable AI: From black box to glass box." *Journal of the Academy of Marketing Science* 48 (2020): 137-141. • Mourtzis, Dimitris, John Angelopoulos, and Nikos Panopoulos. "The future of the human–machine interface (HMI) in society 5.0." *Future Internet* 15.5 (2023): 162. • Sosa, Rosario Uceda, et al. "Reasoning about concepts with LLMs: Inconsistencies abound." *First Conference on Language Modeling*. 2024.	

Tabelle 36: Transparenz und Erklärbarkeit

G-03 - Manipulation, Desinformation und Social Engineering

Risikoübersicht	
Risikokategorie	Gesellschaftliches Risiko
Risiko-ID	G-03
Risikobeschreibung	Künstliche Intelligenz und die damit verbundene Fähigkeit Inhalte jeglicher Art zu erstellen können zu einer zunehmenden Bedrohung für eine offene Gesellschaft und Organisationen zugleich werden. Einerseits gibt es allgemeine Desinformation und Manipulation der breiten Bevölkerung, was besonders für demokratische Prozesse und eine Spaltung der Gesellschaft gefährlich sein kann. Andererseits kann mit denselben Methoden auch Social Engineering im klassischen Sinn betrieben, und ein gezielter Angriff auf einzelne Organisationen gestartet werden. In beiden Fällen kommt es jedenfalls zu einer nachhaltigen Schädigung und qualitativen Abwertung des gesamten Informationsökosystems.
Risikoindikatoren	
Anomalien und unnatürliche Inhalte	Besonders bei von KI generierten Bildern kommt es immer wieder zu bestimmten Elementen, die unnatürlich sind und dadurch ein konkretes Anzeichen, dass der vorliegende Inhalt nicht echt ist. Gleichermaßen ist das Gegenteil, übermäßige Perfektion ohne Anzeichen menschlicher Fehler, ebenfalls ein Anzeichen für generierte Inhalte.
Problematik der Quellen	Von KI generierte Inhalte sind oftmals ohne Quellenangabe, und selbst wenn Quellen angegeben werden, sind diese Angaben schwer verifizierbar oder widersprüchlich.
Polarisierende Inhalte	Wenn Texte, Bilder oder Videos den Zweck der Verbreitung und Aufwiegelung von Emotionen haben, und die vermittelten Inhalte stark polarisierend sind, so kann man von gezielter Manipulation ausgehen. Besonders wenn auch noch zeitliche Dringlichkeit vermittelt wird, um so noch mehr psychologischen Druck auszuüben.
Risikoszenario	
Kampagne zur Schädigung demokratischer Institutionen	Ein staatlicher bzw. staatsnaher Akteur eines autokratischen Landes nutzt generative KI-Systeme zur Erstellung von manipulativen Inhalten als großangelegte Kampagne mit dem Ziel der Vertrauensschädigung demokratischer Institutionen, Polarisierung der politischen Landschaft, und einer damit verbundenen Belastung der nationalen Sicherheitsbehörden. In einem ersten Schritt wird KI dazu verwendet, soziale Medien zu analysieren und relevante Themengebiete zu finden, die dann in weiteren Schritten durch zugeschnittene Desinformation ausgenutzt werden. Für mehr Wirksamkeit werden Fake-Profile und vermeintliche Experten simuliert, um die Stimmung weiter anzuheizen.
Risikokomponenten	
Bedrohung	Manipulation der öffentlichen Meinung, Social Engineering und Cyberangriffe, Vertrauensverlust in Institutionen
Asset (Ursache)	Fortschrittliche generative KI-Modelle und deren Inhalte, Botnetze, Niedrige Bewusstseins- und Aufklärungsrate der Bevölkerung
Asset (Auswirkung)	Konflikte und gesellschaftliche Spaltung, Schwächung der Demokratie, Störung der Märkte
Sicherheitsmaßnahmen	

Psychologische Sicherheit	Eine wichtige Schutzmaßnahme gegen erfolgreiche Desinformation und Manipulation ist die Stärkung der "psychologischen Sicherheit" betroffener Personen und der Gesellschaft als Ganzes. Bspw. Sind die psychologischen Phänomene Confirmation Bias und Social Proof große Treiber, wenn es darum geht falsche Informationen zu akzeptieren.
Erklärbarkeit von KI	Durch eine Stärkung der Erklärbarkeit von KI-Systemen und eine damit verbundene Nachvollziehbarkeit der erstellten Inhalte ermöglichen es, bei eigener Anwendung von KI in der Lage zu sein, Falschinformationen zu identifizieren.
Technische KI-Erkennung	Sowohl von KI generierte Texte wie auch Bild- und Videoaufnahmen können mit technischen Hilfsmitteln erkannt werden. Beispielsweise kann eine linguistische Analyse von Texten Aufschluss darüber geben, ob es sich um einen generierten Inhalt handelt oder nicht. Ebenso können für den Menschen nicht sichtbare Wasserzeichen als Merkmale verwendet werden, um von KI-generierte Inhalte zu erkennen.
Weiterführende Informationen	
• OECD Artificial Intelligence Papers – Assessing potential future artificial intelligence risks, benefits and policy imperatives, November 2024 • Monteith, Scott, et al. "Artificial intelligence and increasing misinformation." *The British Journal of Psychiatry* 224.2 (2024): 33-35. • Aïmeur, Esma, Sabrine Amri, and Gilles Brassard. "Fake news, disinformation and misinformation in social media: a review." *Social Network Analysis and Mining* 13.1 (2023): 30. • Otterbacher, Jahna. "Why technical solutions for detecting AI-generated content in research and education are insufficient." *Patterns* 4.7 (2023). • Jiang, Zhengyuan, et al. "Watermark-based Detection and Attribution of AI-Generated Content." *arXiv preprint arXiv:2404.04254* (2024).	

Tabelle 37: Manipulation, Desinformation und Social Engineering

G-04 - Falsche Erwartungen an KI

Risikoübersicht	
Risikokategorie	Gesellschaftliches Risiko
Risiko-ID	G-04
Risikobeschreibung	Die gesellschaftlichen Erwartungen an KI sind, stark beeinflusst durch mediale Darstellungen und übermäßige Versprechungen durch "Hype", stark verzerrt, und spiegeln nicht die Tatsachen der Realität wider. Besonders Menschen ohne entsprechendes technisches Fachwissen neigen oft dazu, bei KI an fantastische Konstrukte aus Spielfilmen zu denken. Dies führt dann zur großen Enttäuschung, mit der Erkenntnis, dass KI doch "nur" ganz bestimmte Aufgaben ausführen kann, und kein universales Tool mit beinahe magischen Kräften ist. Diese falsche Erwartungshaltung kann besonders bei Entscheidungsträgern zu verschwenderischen Investitionen und falschem Vertrauen führen.
Risikoindikatoren	
Übermäßiges Vertrauen in die Fähigkeiten von KI	Auch ein KI-System muss, wie jedes andere System, gewartet und überwacht werden. Die Fähigkeiten eines KI-Systems können jedoch dazu verleiten, dass dem System mehr Vertrauen zugeschrieben wird als es sein sollte. Beispielsweise kann es so zum falschen Einsatz eines KI-Systems kommen, obwohl das System nicht für jenen konkreten Anwendungsfall ausgelegt ist.
Schuldzuweisungen an das KI-System	Ein äußerst problematisches Anzeichen sind Schuldzuweisungen an ein KI-System. Das Vertrauen in die Fähigkeiten von KI ist dermaßen hoch, dass Fehler nicht mehr erwartet werden. Kommt es dennoch wieder zu Fehlern, muss es die Schuld der KI gewesen sein, und nicht die eigene falsche Anwendung oder mangelnde Übersicht.
Blindes Vertrauen	LLMs werden immer häufiger als ständiger Begleiter verwendet. Oftmals kommt es dabei vor, dass Endanwender die Ausgaben eines solchen KI-Systems, ohne es zu hinterfragen, als wahr annehmen. Dieses blinde Vertrauen kann zu falschen Handlungen und Entscheidungen der Menschen führen.
Risikoszenario	
Selbstfahrende Autos	Beim Fahren auf der Autobahn mit einem selbstfahrenden Fahrzeug kommt es kurz vor dem Beginn einer Baustelle zu sich spontan verändernden Licht- und Wetterverhältnisse. Die Sensoren und Kameras sind durch diese Verschlechterung der Fahrverhältnisse und stark reflektierendes Licht nicht mehr in der Lage eine korrekte Entscheidung zu treffen. Dadurch erkennt das Fahrzeug nicht, dass eine Spurumleitung bevorsteht. Auch Warnsignale der anwesenden Arbeitskräfte werden nicht erkannt, da diese nicht Teil des Trainingssets waren. Somit fährt das Auto, im Glauben, dass die Straße normal weiterführt, direkt in die Baustelle ein und es kommt zu einem Verkehrsunfall bei hoher Geschwindigkeit, bei dem Personen durch volles Vertrauen in ein KI-System zum Schaden kommen.
Risikokomponenten	
Bedrohung	Automation Bias, Fehlanwendung von KI
Asset (Ursache)	Mangelnde Aufklärung und Bildung, Marketing-Hype und Medien, mangelnde Transparenz

Asset (Auswirkung)	Schaden für Menschen und Wirtschaft, Vertrauensverlust in die Technologie, Fehlentscheidungen des Menschen
Sicherheitsmaßnahmen	
Schulungen und Bewusstsein	Durch entsprechende Aufklärungsmaßnahmen können Fähigkeiten, Verantwortung, und Erwartungshaltungen gegenüber KI-Systemen kommuniziert werden, sodass Endanwender und Entscheidungsträger besser auf den Umgang mit KI vorbereitet sind.
Einforderung von Verantwortlichkeit	Wenn Endanwendern klar gemacht wird, dass sie für die Verwendung von KI-Ergebnissen gleichermaßen Verantwortung tragen, wie das bei herkömmlichen Ansätzen ohne KI bisher der Fall gewesen ist, kann ein höheres Maß an Verantwortungsbewusstsein und Ergebnisüberprüfung bezweckt werden.
Explainable AI	Erklärbarkeit und ganz besonders ein Fokus auf herausragende Mensch-Maschine-Interaktionen kann dabei helfen, dass Endanwender nicht nur ein KI-System und seine Fähigkeiten, sondern auch interne Abläufe und damit verbundene Ergebnisse besser verstehen.
Weiterführende Informationen	
• Lyell, David, and Enrico Coiera. "Automation bias and verification complexity: a systematic review." *Journal of the American Medical Informatics Association* 24.2 (2017): 423-431. • Goddard, Kate, Abdul Roudsari, and Jeremy C. Wyatt. "Automation bias: a systematic review of frequency, effect mediators, and mitigators." *Journal of the American Medical Informatics Association* 19.1 (2012): 121-127. • Nordhoff, Sina, and Joost De Winter. "Why do drivers and automation disengage the automation? Results from a study among Tesla users." *arXiv preprint arXiv:2309.10440* (2023). • SOH, Jerrold. "The executive's guide to getting AI wrong." (2022): 74. • Awad, Edmond, et al. "Drivers are blamed more than their automated cars when both make mistakes." *Nature human behaviour* 4.2 (2020): 134-143. • Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." *Artificial intelligence* 267 (2019): 1-38. • Dwivedi, Rudresh, et al. "Explainable AI (XAI): Core ideas, techniques, and solutions." *ACM Computing Surveys* 55.9 (2023): 1-33.	

Tabelle 38: Falsche Erwartungen an KI

Explorative Szenarioanalyse – Methodische Aspekte

Das Thema KI in der LEA ist äußerst komplex, denn die rasante Entwicklung neuer Technologien und spezifischer Anwendungen macht es erforderlich, die gewonnenen Erkenntnisse regelmäßig zu überprüfen, um ihre Gültigkeit zu erhalten. Aus diesem Grund haben wir uns in der Studie auf zwei Schlüsselstrategien konzentriert, um sicherzustellen, dass die Ergebnisse für die Entscheidungsträger so nützlich wie möglich sind. (1) Reproduzierbarkeit und (2) Erkundung künftiger Entwicklungen. Bei der Exploration werden Methoden der Explorativen Szenarioanalyse (ESA) eingesetzt, um zukünftige Entwicklungen auf der Grundlage technologischer, rechtlicher und gesellschaftlicher Dimensionen zu extrapolieren. Diese Extrapolationen erzeugen Szenarien, die dann zur Untersuchung der daraus resultierenden Herausforderungen genutzt werden können. Diese Lücken können später als wertvoller strategischer Input dienen.

In diesem Kapitel werden wir die Methodik der Explorativen Szenarienanalyse beschreiben, dabei bauen wir auf einem sehr bekannten Ansatz von Geschka et al. auf, den wir bereits in einem kooperativen KIRAS-Projekt (exploreAI) genutzt hatten. Allerdings hatte dieser Ansatz einige Lücken und Unstimmigkeiten ergeben, wenn er im technischen Umfeld angewendet wurde. Dies betrifft vor allem die Faktoren, sowie die Faktorausprägungen, aber auch Themen wie das Scoring und die Nutzung rein proprietärer Software und Lösungen.

Außerdem war der Ansatz von einer sehr hohen Komplexität geprägt, die ihn durch Nicht-Expert*Innen mit gutem sozialwissenschaftlichem Background kaum anwendbar machte. Auch war durch das Scoring bedingt die Diversität der ausgewählten Szenarien gering, die Methodik behalf sich dabei mit dem Einsatz willkürlich ausgewählter „Trendbruchereignisse“, die allerdings vollkommen aus der recht starren und stringenten Methodik ausbrachen und mehr oder weniger die Inklusion von verschiedenstem Expert*Innenwissens, aber auch persönlichen Ansichten, ermöglichte. Dies führt unserer Meinung nach dazu, dass die Methodik auf der einen Seite den Aufwand einer extrem exakten und mathematisch durchkonstruierten Methodik aufwies, gleichzeitig aber durch die Trendbruchereignisse und den recht frei vergebenen Basiswerten für die Scorings eine Art vorgetäuschte Genauigkeit zeitigte, bei der sich finale Scores von Szenarien in der 2 Nachkommastelle unterschieden, die Eingangsvariablen aber Pi mal Daumen mit hoher Unschärfe gewählt wurden. Auch kamen damals sehr gleichförmige Szenarien heraus, die wenig Diversität besaßen, was ja eigentlich dem Hintergedanken der Planung und Vorbereitung auf verschiedene, aber wahrscheinliche, Szenarien zuwiderläuft.

Dabei möchten wir festhalten, dass dies keine generelle Kritik an der Methode von Geschka ist, diese ist nicht für diese Art von Fragen konzipiert worden und daher durch die Kritik auch nicht direkt betroffen, sie eignete sich unserer Meinung nach lediglich für diese Analysegebiet nur sehr eingeschränkt.

Grundlagen und Basisansatz nach Geschka

Wir basierten unsere Explorative Szenarienanalyse stark auf dem Ansatz von Geschka[1] und dessen Weiterentwicklungen[2], den wir für unsere Zwecke modifizierten, wobei wir nicht nur den Ablauf veränderten, sondern auch die Begrifflichkeiten anders verwendeten, um sie mehr an unsere Systematiken anzupassen. Grundsätzlich ist der Ablauf aber sehr ähnlich, für ein Untersuchungsfeld wird analysiert, welche Einflussfaktoren und wesentlich sind, dann werden deren Ausprägungen analysiert, eine konsistente Auswahl wird zur Szenariengenerierung herangezogen, wobei Geschka et al. noch sog. Trendbruchereignisse einführen, die wir in unserer Methodik aufgrund unserer Anpassungen weniger beachten müssen, die aber durchaus wertvolle Eigenschaften zur Modellierung beitragen. Abbildung 39 zeigt einen Überblick über den originalen Ablauf.

[1] Geschka H. (2006) Szenariotechnik als Instrument der Frühaufklärung. In: Gassmann O., Kobe C. (eds) Management von Innovation und Risiko. Springer.
[2] Von Reibnitz, Ute (2013): Szenario-Technik: Instrumente für die unternehmerische und persönliche Erfolgsplanung. Springer.

Abbildung 39: Explorative Szenarienanalyse nach Geschka

Definition und Strukturierung des Untersuchungsfeldes: In der ersten Phase wird der Untersuchungsgegenstand genau spezifiziert, vor allem in Hinblick auf eine Abgrenzung. Dies ist besonders wichtig, um die Analysen nicht ausufern zu lassen und einen klaren Fokus zu gewährleisten. Dazu ist es wesentlich, dass die relevanten Hintergrundinformationen gesammelt und zur Verfügung gestellt werden. In dieser Phase wird auch der Zeithorizont der Explorativen Szenarienanalyse festgelegt, meist auf Basis der Analyse eben dieser Hintergrundinformationen.

Identifikation der wichtigsten Einflussfaktoren und Einflussbereiche: Diese Phase besteht aus vier Schritten: (i) Die Sammlung von Einflussfaktoren, (ii) Aufstellen einer Wirkungsmatrix zur Abbildung sich gegenseitig beeinflussender Einflussfaktoren, (iii) Berechnung der kumulierten Einflussstärke, sowie (iv) Erstellung und Betrachtung des System Grids. Dabei ist zu beachten, dass der Terminus „Einflussfaktor“ hier wesentlich abstrakter ist, als dies in der weiteren Analyse basierend auf der A3-Methodik für die Faktoren der Fall ist, da in der nächsten Phase noch die Deskriptoren hinzukommen, die eher den Faktoren in der A3-Methodik entsprechen.

- **Sammlung der Einflussfaktoren**: Es werden mehrere Möglichkeiten der Sammlung vorgeschlagen, meist basierend auf der akademischen Literatur, aber auch in Gruppen durch Kreativtechniken wie der Kartenumlauftechnik. Bei den Einflussfaktoren handelt es sich um Kräfte, die sich auf das Untersuchungsfeld auswirken. Sollten sehr viele Einflussfaktoren gesammelt worden sein, so können diese auch noch in sog. Einflussbereiche zusammengefasst werden.
- **Wirkungsmatrix**: Diese ähnelt sehr stark der Wirkungs- und Konsistenzmatrix im A3-Ansatz, allerdings arbeitet sie auf Ebene der Einflussfaktoren, also 2-3 Ebenen über den Faktorausprägungen (die Deskriptoren der folgenden Phase entsprechen eher den Faktoren in A3), Zur Erstellung wird eine Matrix aufgebaut, in der in jeder Zeile ein Einflussfaktor steht. Der gleiche Einflussfaktor steht auch in der entsprechenden Spalte, sodass eine symmetrische Matrix entsteht, in deren erster Zeile und erster spalte die Einflussfaktoren stehen, der Rest ist leer. Die Diagonale wird ausgegraut, da sie die gleichen Einflussfaktoren in Zeile und Spalte enthält.
- **Berechnung der kumulierten Einflussstärke**: Die Wirkungsmatrix wird nun mit Werten aus $\{0, 1, 2\}$ befüllt, wobei „0“ keinen Einfluss, „1“ schwachen Einfluss und „2“ starken Einfluss bedeutet. Dabei bedeutet ein „1“er in der zweiten Zeile und der ersten Spalte, dass der Einflussfaktor in der zweiten Zeile schwachen Einfluss, auf den in der ersten Spalte besitzt. Nun kann Zeilen- und spaltenweise aufsummiert werden, man erhält dadurch zeilenweise gelesen die Stärke, mit der der Einflussfaktor der zweiten Zeile das Untersuchungsfeld, aufgespannt durch die Menge der Einflussfaktoren, beeinflusst. Umgekehrt erhält man durch die Spaltensumme, wie stark das Untersuchungsfeld den jeweiligen Einflussfaktor beeinflusst.
- **Aufstellen des System Grids**: Das System-Grid ist eine Darstellung der Einflussfaktoren in einem einfachen Koordinatensystem, wobei die kumulierte Einflussstärke und die kumulierte Beeinflussbarkeit die Achsen ausbilden. Durch Aufteilen der eingetragenen Einflussfaktoren in vier Quadranten wird zwischen aktiven, ambivalenten, puffernden und passiven Einflussfaktoren unterschieden. Dieses Grid wird für die Auswahl der weiter zu nutzenden Einflussfaktoren genutzt, je nachdem ob mehr aktive oder mehr passive Einflussfaktoren in die Analyse einbezogen werden sollen.

Formulierung von Deskriptoren und Aufstellung von Projektionen und Annahmen: Dieser Schritt bricht die Einflussfaktoren in konkretere Begriffe herunter, d.h. die Falktoren werden in Unterthemen aufgebrochen. Beispiel: der Faktor „Arbeitsmarkt“ wird in die Deskriptoren „Arbeitslosenquote“, „Teilzeitquote“, „Saisonarbeit“ … . aufgeteilt. Diese Deskriptoren entsprechen eigentlich den Faktoren in der A3-ESA. Diese Phase besteht aus den folgenden Teilschritten:

- ***Finden von Deskriptoren***: Diese Begriffe müssen durch ein Einzelergebnis vollständig beschrieben werden können. Einflussfaktoren haben in dieser Methode umso mehr Einfluss auf die Szenariengenerierung, je mehr Deskriptoren ihnen zugeordnet werden können.
- ***Definition IST-Zustand***: bevor Projektionen in de Zukunft durchgeführt werden können, muss der tatsächliche Zustand zum gegenwärtigen Zeitpunkt definiert werden.
- ***Aufstellen von Projektionen und Annahmen***: Hierbei kann auf das Wissen von expert*innen zurückgegriffen werden, um mögliche Annahmen für die Zukunft zu beschreiben. Dabei gibt es eindeutige Deskriptoren, die nur eine sinnvolle Entwicklung zulassen, sowie solche mit mehreren Entwicklungsmöglichkeiten, die Alternativ- Deskriptoren. Ein Deskriptor gilt als vollständig beschrieben, wenn vollständiger Ereignisraum bestehend aus allen erdenklichen Ereignissen abgebildet ist.

Bildung und Auswahl alternativer konsistenter Annahmekombinationen: In diesem Schritt werden die Faktorausprägungen gruppiert nach den zugehörigen Deskriptoren in eine sog. Konsistenzmatrix eigetragen, d.h. eine Matrix ähnlich der Beeinflussungsmatrix, lediglich repräsentieren die Werte die Konsistenz. Der genaue Wertebereich hängt von der verwendeten Software ab und ist typischerweise nicht transparent.

Entwicklung und Interpretation der ausgewählten Szenarien: Wieder wird durch proprietäre Software die Menge aller konsistenten Faktorausprägungen berechnet, wobei auch Werte berechnet werden, wie gut die einzelnen Faktorausprägungen zu dem Faktorausprägungsbündel passen und damit ein Score für das gesamte Bündel berechnet. Diese Bündel bilden die Ausgangspunkte für die fertig zu beschreibenden Szenarien. Außerdem wird zu jedem ausgewählten Bündel auch das dazu duale Kontrastbündel als Szenario-Nukleus berechnet.

Einführung und Analyse signifikanter Trendbruchereignisse: Das bisherige Vorgehen kann nicht gut mit plötzlichen, disruptiven Ereignissen umgehen, die die Entwicklungsrichtung empfindlich stören können, den sog, Trendbruchereignissen. In diesem Schritt wird untersucht, welche Trendbruchereignisse die bisher definierten Szenarien-Nuklei beeinflussen könnten, wobei dies durch verschiedenste Kreativtechniken geschehen kann. Daraus können neue Varianten der Szenarien, oder aber Präventivmaßnahmen definiert werden.

Ausarbeitung der Szenarien und Ableitung von Konsequenzen: Die Szenarien werden in Form von längerer Prosa ausgearbeitet[3], wobei weitere kreative Ausgestaltung mit gewissen Abweichungen erlaubt ist. Darauf aufbauend wir eruiert, welchen Einfluss die Szenarien auf die ursprüngliche Fragestellung und das Untersuchungsfeld haben und welche Konsequenzen aus diesen Erkenntnissen zu ziehen sind. Dies kann sich speziell auch auf mögliche Gefahren beziehen.

Konzipieren von Maßnahmen und Planungen: Auf Basis der vorangegangenen Erkenntnisse können Maßnahmen abgeleitet werden, diese entstehen jedoch nicht direkt aus den Szenarien selbst, die Strategieentwicklung ist in sich ein kreativer Prozess, der neben den Erkenntnissen aus

[3] Mietzner, D. (2009): Strategische Vorausschau und Szenarioanalysen: Methodenevaluation und neue Ansätze. Springer.

der ESA einiges an Domänenwissen erfordert. Für die weitere Durchführung stehen viele Techniken aus der Strategieentwicklung zur Verfügung.

Methodische Kritik

Schon bei der ersten Durchführung der Explorativen Szenarienanalyse nach Geschka wurden die Nachteile für die wiederholte Nutzung im technischen Umfeld erkannt und entsprechend schon in der ersten Phase der Durchführung dieser Studie behoben. Im Rahmen der weiteren Analysen im Zuge dieser Studie, speziell auch in der ersten Durchführung der Fragestellungen, kamen weitere Kritikpunkte hinzu. Folgend eine Übersicht über die Kritikpunkte:

- **Generierung der Faktoren**: In der originalen Methodologie war dies nicht Teil von Interviews, sondern basierte auf Einschätzungen aus der Literatur. Dies ist in den ursprünglich vorgeschlagenen Anwendungsgebieten sicherlich sehr gut durchführbar, da hier auf langjährige Literatur verwiesen werden kann. In einem Feld mit einer derartig schnellen Entwicklung wie das der AI liegt das Problem darin, dass die akademische Literatur um Monate, wenn nicht Jahre hinterherhinkt, d.h. die Möglichkeiten, die mit dem sehr kurzfristigen Einsatz sehr fundamentaler AI-Möglichkeiten, erreichbar für normalen Bürger*innen, entstanden, werden erst in der künftigen Literatur entsprechend rezipiert. Dies ist auch in Bezug auf Security-Analysen ein Problem, da auch hier immer wieder Ereignisse stattfinden, die die Betrachtung eines Gegenstandes grundsätzlich verändern und komplett neue Faktoren ins Spiel bringen, wie bspw. grundlegende Sicherheitsprobleme auf Hardware-Seite und dergleichen. Diese reduzieren den Wert bisheriger Annahmen teilweise deutlich. Auch Änderungen im technologischen Umfeld, wie bspw. das Auflassen von SGX im Bereich des Trusted Computings, führen zu starken Trendbrüchen. Daher war es notwendig, die wesentlichen Faktoren aus den Interviews mit Expert*innen zu extrahieren.
- **Beeinflussung durch die ExpertInnen-Szenarien**: Durch diese Änderung ergab sich eine weitere Problematik, die auch im originalen Ansatz beachtet werden muss. Auch wenn die Faktoren direkt aus der Literatur gewonnen werden, so ist es dennoch oft nicht möglich, diese von den Ansichten der Autor*innen zu trennen. Dies bedeutet, dass bei einem Interview, aber auch in Artikeln, zu künftigen Technologieentwicklungen auch immer die Ansicht der Autor*innen zur tatsächlichen Gestalt dieser Entwicklungen mit einfließen, d.h. man bekommt nur selten neutrale Einflussfaktoren genannt, sondern immer auch eine Ansicht zu tatsächlich möglichen und wahrscheinlichen Ausprägungen, d.h. man bekommt eher ein Set von Faktorausprägungen statt Faktoren, dennoch müsste man nach dem Konzept alle Faktorausprägungen für einen Faktor gleichmäßig betrachten.
- **Extreme Breite der generischen Faktoren**: In den typischen Beispielen aus der Literatur, aber auch in den ersten Durchführungen durch uns selbst, wurden wir der extremen Breite der Faktoren gewahr, die dazu führt, dass, wenn eine gute Abdeckung aller möglichen Faktorausprägungen abgedeckt werden soll, nur sehr wenige Faktoren für die Generierung der Szenarien ausgewählt werden können (siehe nächster Punkt). Dabei stellt sich die Frage, ob die Faktoren in dieser Breite (i) überhaupt wahrscheinlich, (ii) überhaupt von Interesse in der Szenariengenerierung und (iii) nicht zu ähnlich in Bezug auf die Effekte der generierten Szenarien auf den Untersuchungsgegenstand sind.
- **Geringe Menge an ausgewählten Faktoren**: Bedingt durch die große Breite der Faktoren werden in den uns bekannten Beispielen für die Nutzung der ESA nur sehr wenige Faktoren ausgewählt, meist 3-5. Dies führt dazu, dass nur sehr wenige extrapolierte Faktoren tatsächlich einfließen. Auch hierbei möchten wir gerne festhalten, dass dies in den

ursprünglichen Anwendungsgebieten sinnvoll und aussagekräftig sein dürfte, im Bereich der Abschätzung von Technologieentwicklungen ist uns diese Menge an Faktor(ausprägungen) allerdings zu klein, da viele verschiedene Effekte (siehe auch die tatsächliche Auswertung) wesentlich für die weitere Entwicklung sein können, die nicht direkt voneinander abhängen.

- **Konsistenzprüfung auf Faktorebene**: Im originalen Ansatz wird die Konsistenzprüfung auf einer Ebene ausgeführt, die den Faktorausprägungen entspricht, allerdings sind die verwendeten Scoring-Mechanismen komplett intransparent und proprietär, was die Nachvollziehbare t hemmt. Gleichzeitig gibt es verschiedene Scorings, je nach Softwareprodukt, d.h. die Methodologie spricht nur von einem Scoring, aber ohne dieses genau zu definieren.
- **Scoring der Relevanz auf Faktorebene**: Auch die Relevanz wird auf Ebene der Faktoren und nicht der Faktorausprägungen eruiert, basierend darauf, welche Faktoren, welche anderen wie stark beeinflussen. Auch hierbei sind wir der Meinung, dass in vielen technischen Gebieten die Stärke (oder überhaupt das Vorhandensein) der Beeinflussung wesentlich von der Faktorausprägung abhängt. Beispiel: Faktor „Nutzung von AI“ (Ausprägungen: Allgegenwärtig – Für viele Aufgaben – Für ausgewählte Aufgaben – Nie) und Faktor „Energiepreise“ (Ausprägungen: Sehr niedrig – Mittel – Sehr hoch). Der Faktor „Energiepreise“ wird hohen Einfluss auf bspw. Faktorausprägung „Allgegenwärtig“ haben, da sehr hohe Energiepreise diese Faktorausprägung eher unmöglich machen werden, gleichzeitig kann der Effekt der Energiepreise auf „Nie“ auch recht niedrig eingestuft werden, vor allem wenn diese „Niedrig“ sind, die Akzeptanz von AI aber durch andere Faktoren wesentlich bestimmt werden. Auch hierzu sollte die Relevanz auf Ebene der Faktorausprägungen und nicht der Faktoren bestimmt werden.
- **Dominante triviale Faktoren und Faktorausprägungen**: Wir haben in der Abschätzung der Wichtigkeit einzelner Faktoren bewusst die Komplexität des Scorings der Relevanz der Faktoren reduziert, da gewisse Faktoren und Faktorausprägungen extrem dominant, aber eigentlich uninteressant sind, gleichzeitig aber durch die starke Abhängigkeit vieler anderer Faktoren von diesen wenigen Faktoren und Faktorausprägungen, in der Szenariengenerierung dominieren. Ein Beispiel wäre „Energiepreise“, die zwar offensichtlich wichtig sind, aber bei einer Analyse in Bezug auf LEA-spezifische Fragestellungen zu Ai nicht den Kern der Fragestellung treffen. Wir wissen, dass günstige Energie wichtig für die AI-Nutzung ist, dies ist aber eher nicht die Fragestellung einer LEA. Dabei spielt auch eine Rolle, dass die tatsächliche Eintrittswahrscheinlichkeit einer Faktorausprägung nicht in das Scoring mit einfließt, dieses im Wesentlichen darauf beruht, wie viele andere Faktoren durch den gegebenen Faktor beeinflusst werden.
- **Proprietäres Scoring**: Das im Rahmen des originalen Ansatzes durchgeführte Scoring der Szenarien ist proprietär und benötigt die Nutzung einer speziellen Software (INKA). Die bedeutet auch, dass die Nachvollziehbarkeit der Generierung nicht wirklich gegeben ist, zwar ist das Scoring deterministisch, aber (i) kann es nicht transparent nachvollzogen werden und (ii) ist es unklar, wie stabil der Mechanismus in Bezug auf kleine Änderungen in der Abschätzung von Wichtigkeit der einzelnen Faktoren ist. Es ist auch unklar, ob durch das Hinzufügen von in sich trivialen und unwichtigen zusätzlichen Faktoren, die aber von einem bestimmten Faktor stark abhängen, der Score dieses Faktors in die Höhe getrieben werden, die Auswahl von an sich unwichtigen und nichtberücksichtigten Faktoren also auch das Scoring beeinflussen kann. Da wir hierzu kein Paper fanden und die Software proprietär ist, ist eine Aussage zum originalen Scoring für uns nicht möglich.

- **Potenziell hohe Ähnlichkeit der Szenarien**: Speziell im Fall von sehr stark dominierenden Faktoren/Faktorausprägungen müsste das originale Scroing dennoch Szenarien basierend auf sehr ähnlichen Faktoren den Vorzug geben, d.h. es ist nicht transparent, inwieweit eine gewisse Diversität in der Szenarienerstellung erreicht werden kann. Gerade diese ist aber spannend, um zu identifizieren, welche Anforderungen und Probleme in möglichst vielen wahrscheinlichen Szenarien auftreten und entsprechend diese Gaps zu priorisieren.
- **Duale Szenarien**: Nutzende der ESA nach Geschka wählen oft für jedes generierte Szenario das duale Szenario als zusätzliche Vorhersage aus, indem die einzelnen Faktoren mit möglichst gegenteiligen Faktorausprägungen besetzt werden, dennoch aber konsistent sein sollen. Dies ist zwar grundsätzlich spannend, da aber in manchen Falktoren zwischen allen befragten Expert*innen große Einigkeit bezügliche potenziell sinnvoller Faktorausprägungen herrschte, hätte dies die Nutzung von Faktorausprägungen bedeutet, denen kaum Relevanz zugebilligt worden ist – als gleichwertiges Szenario. Wir verzichten daher vollständig auf das duale Szenario, da es quasi eine Art „Anti-Trend“ zu den Expert*innen-Meinungen darstellen würde.
- **Potenziell vorgetäuschte Genauigkeit der Scores**: Durch die rigorose Art des Scorings kommt es zu einer Methode, die sehr stringent ist und die Auswahl der Faktorausprägungen und Faktoren methodisch sehr stark begründet. Allerdings basieren die Bewertungen wiederum auf recht ungenauen und fehleranfälligen Methoden, wie der manuellen Auswahl aus der akademischen Literatur und der Meinung von expert*innen. Daher muss sehr darauf geachtet werden, dass es beim Scoring nicht zu einer vorgetäuschten Genauigkeit kommt, d.h. die Berechnung sehr genauer Scores und die Nutzung sehr spezifischer Threshholds, obwohl die Genauigkeit der Ursprungsdaten dies überhaupt nicht hergibt. Dies ist vor allem ein Problem, wenn sich durch Umbrüche langjährig etablierte wichtige Faktoren und Faktorausprägungen ändern, bzw. deren Gewicht im Feld.
- **Trendbruchereignisse als methodischer Bruch**: Die originale Methode von Geschka ist sehr komplex und stark methodisch geschlossen und zielt darauf ab, die Szenarien sehr stark von einem Score abhängig zu machen. Da aber auch in den ursprünglichen Anwendungsgebieten immer wieder sehr stark durchschlagende disruptive Änderungen passieren, kann all dies durch die Einführung von Trendbruchereignissen überstimmt werden. Dies ist ein methodischer Bruch, da alle vorangegangenen stringenten Berechnungen quasi außer Kraft gesetzt und durch eine neue Faktorausprägung oder gar ein (partielles) Szenario ersetzt werden.

Die vorangegangene methodische Kritik führte zu einer Vielzahl an Änderungen am Vorgehen der Explorativen Szenarioanalyse, die eine deutliche Vereinfachung versprechen und damit auch die wiederholte Durchführbarkeit viel leichter ermöglichen. Dies führt dazu, dass nicht eine ESA für viele Jahre als Referenz herangezogen werden muss, sondern die Methode mit immer neuen Expert*Innen, neuen Use-Cases und neuen Daten und Informationen im Schritt der Entwicklung iterativ durchgeführt werden kann, ohne den Aufwand eines jeweils neuen Forschungsprojekts mit den entsprechenden Ausgaben zu zeitigen. Wir möchten hierbei noch einmal hinweisen, dass sich die vorgebrachte Kritik nicht an die Methodik der ESA nach Geschka selbst richtet, sondern sehr stark durch die Anwendung in technologischen Gebieten mit stark disruptivem Charakter bedingt ist.

Finale Methodik der ESA nach A3

In diesem Kapitel stellen wir die neu überarbeitete Version de Explorativen Szenarienanalyse nach A3 vor. Dabei wurde versucht, auf der einen Seite den Kritikpunkten im vorangegangenen Abschnitt gerecht zu werden und die Methodik entsprechend anzupassen, auf der anderen Seite sollte die neue Methodik wesentlich einfacher anwendbar und wiederholbar sein.

Wir sind der Meinung, dass uns dies gelungen ist, dabei ist anzuführen, dass eine ESA nie eine exakte und eindeutige Folgerung aus Werten sein kann: Alleine schon die Auswahl der Faktoren im Rahmen der Recherche beeinflusst das Ergebnis nachhaltig, auch andere Aspekte, wie die Bewertung von Faktoren und Faktorausprägungen sind stark subjektiv gefärbt. Eine zu verwissenschaftlichte und mit komplexen Scores hinterlegte Methode birgt unserer Meinung nach die Gefahr einer vorgetäuschten Exaktheit, die alleine wenn man die Entscheidungsfindung bezüglich Auswahlen und Scorings betrachtet nicht haltbar ist.

Zusätzlich muss die Zielsetzung einer ESA betrachte werden: Es geht nicht nur darum, möglichst wahrscheinliche Szenarien zu generieren, sondern auch eine Auswahl an verschiedenen Szenarien – dies ist vor allem ein Problem beim Vorhandensein vieler Faktoren mit vielen Faktorausprägungen, sowie dominanter Fakoren, die in die Scorings extrem stark einfließen – und so in jedem Szenario gleich vorhanden sind. Daher haben wir vor allem im Schritt der Auswahl ein wenig mehr Intuition eingebaut und nicht streng nach Score geurteilt, sondern auf Basis der Expert*Inneninterviews als wichtig und prominent erkannte Faktorausprägungen explizit ausgewählt. Dadurch konnten wir auch auf den „Trick“ mit den Trendbruchereignissen aus der Originalmethode verzichten.

Abbildung 40 gibt einen Überblick über die neue Methodik

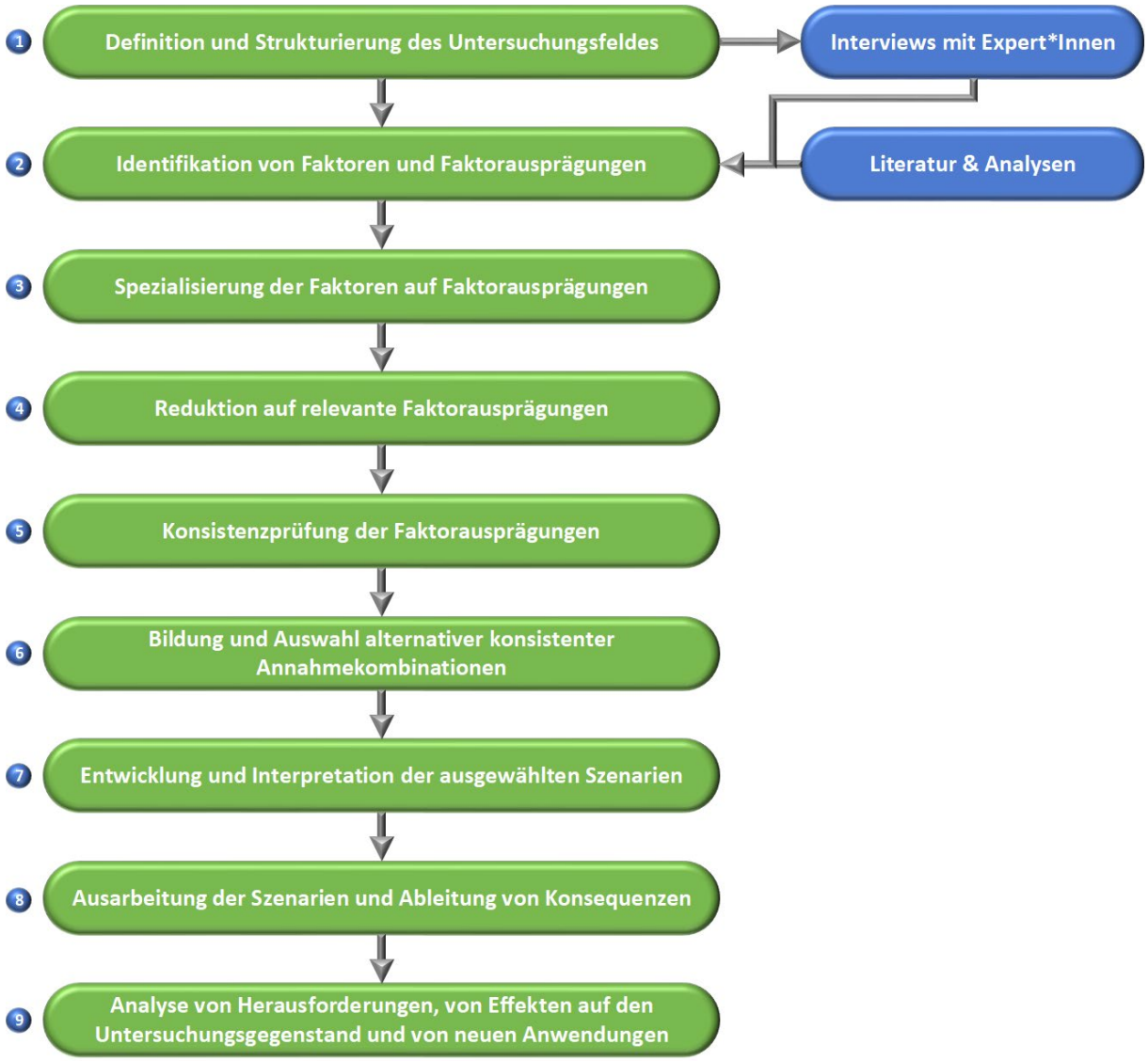


Abbildung 40: Explorative Szenarienanalyse nach A3

Definition und Strukturierung des Untersuchungsfeldes: Dieser Schritt wird weitestgehend analog zum originalen Konzept durchgeführt, allerdings strukturieren wir hier schon den Untersuchungsgegenstand in Dimensionen. Diese dienen der Auswahl der geeigneten Interviewpartner, die bereits in diesem Schritt hinzugezogen werden. Der Hintergrund ist der, dass die Faktoren im originalen Ansatz eher direkt aus der Literatur ermittelt wurden, nicht aus den Interviews. Gerade in einem hochdynamischen Feld wie Ai und auch Cyber-Security, muss davon ausgegangen werden, dass die verfügbaren Publikationen dem aktuellen stand hinterherhinken. Dies betrifft vor allem die derzeitige Situation, wo auf einmal Faktoren relevant wurden, die bisher weniger Beachtung verdient hatten, wie die geopolitische Stabilität. Aber auch technologisch sehen wir im Gebiet der AI kurzfristige riesige Innovationsschritte, die relativ neue Literatur rasch veraltet werden lässt. Auch in Cyber-Security sind wir mit einem Feld befasst, in dem es oft zu sehr sprunghaften Neuerungen und Änderungen kommt, diese aber in der Literatur nicht ausreichend abgebildet sind. Wir sehen daher die Zuziehung von Expert*Innen in diesem Schritt als wesentlichen Erfolgsfaktor für die Fokussierung auf die wesentlichen Faktoren und Faktorausprägungen.

Identifikation von Faktoren und Faktorausprägungen: dieser Schritt ist deutlich komplexer als in der originalen Methodik, da er ihn wesentlich erweitert: Wesentliche Faktoren können natürlich weiterhin aus einer Literaturanalyse, aber auch aus der Analyse von Use-Cases gewonnen werden, wir dies im Original erzielt wird. Dabei fokussieren wir uns in der neuen Methodik allerdings wesentlich mehr auf die möglichen Faktorausprägungen und Scoren diese auch gleichzeitig basierend auf den Aussagen zu ihrer Wichtigkeit in der Literatur. Dabei ist es wesentlich, dass die Literatur sehr kritisch betrachtet wird, da sie speziell in disruptiven technologischen Gebieten der Realität weit hinterherhinken kann und in diesem Fall in Bezug auf ihre Relevanz deutlich geringer eingestuft werden muss. Zusätzlich zur Literatur legen wir daher sehr großen Wert auf die Generierung von Faktoren und Faktorausprägungen aus den Interviews, wobei diese direkt abgefragt werden. Dies sollte möglichst am Anfang der Fragen passieren, damit die Färbung durch den Interviewverlauf möglichst gering ist. Dabei hat es sich für uns bezahlt gemacht, diese wirklich ganz konkret anzufragen, anstatt mit impliziten Fragen zu versuchen herauszufinden, was für den/die Expert*in relevant ist. Aus der Erfahrung bekamen wir in diesen Fragestellungen eher Faktorausprägungen geboten, die zu Faktoren generalisiert werden können. Dies ist insofern positiv, als dies für die Einteilung der Faktoren in sinnvolle und von Expert*innen antizipierte Faktorausprägungen genutzt werden kann, anstatt diese rein aus der akademischen Literatur oder aus dem eignen Wissen, bzw. der eigenen Abschätzung, zu generieren. Zusätzlich können in den Interviews noch mehr Aspekte abgefragt werden, speziell in Bezug auf neue Anwendungen einer (extrapolierten) technologischen Entwicklung und spezifisch nach Herausforderungen und Gefahren. Dies ist speziell sinnvoll, da im Gegensatz zu rein beobachtenden Studien in den Sozialwissenschaften, viele Fragestellungen im technologischen Sektor eine gewisse Wahlfreiheit von Akteur*innen beinhalten, eine neue Technologie, bzw. die Anwendung einer neuen Technologie, zu nutzen oder dies zu verweigern, bzw. eingeschränkt zuzulassen.

Spezialisierung der Faktoren auf Faktorausprägungen: Dies ist ein wesentlicher Bruch mit der originalen Methodologie, anstatt Konsistenz und Relevanz auf der Faktorebene zu beurteilen, wird ab diesem Schritt im Wesentlichen mit Faktorausprägungen gearbeitet, die lediglich gemeinsamen Faktoren zugeordnet werden. Dies reduziert die Probleme der Konsistenz und Beeinflussung auf Faktorebene, wie im vorangegangenen Abschnitt dargelegt. Zusätzlich, und fast noch wesentlicher, erlaubt es die Reduktion der Komplexität: Im originalen Ansatz musste die Auswahl der Faktoren stark reduziert werden, da der durch die verschiedenen Faktorausprägungen aufgespannte Entscheidungsraum sehr groß wurde. In der adaptierten Methodologie können wir durch die Spezialisierung nur mehr die Faktorausprägungen auswählen, die wir für relevant halten und diese auf Konsistenz und Beeinflussung prüfen – wir reduzieren also den Aufwand und erlauben es damit, wesentlich mehr zu unterschiedlichen Faktoren gehörenden Faktorausprägungen für die Szenariengenerierung heranzuziehen. Dies erlaubt es uns auch, Trendbruchereignisse als spezifische Faktorausprägungen von Faktoren zu definieren, denen wir nur wenig Faktorausprägungen zugestehen. Somit wird dieser Bruch in der originalen Methodik aufgehoben. Dabei ist wesentlich, dass diese Auswahl konkret argumentiert wird, bspw. dadurch, dass spezifische Faktorausprägungen von expert*innen genannt wurden.

Konsistenzprüfung der Faktorausprägungen: Dieser Schritt wird grundsätzlich analog zum originalen Ansatz durchgeführt, wobei der frühzeitige Fokus auf Faktorausprägungen zu einigen Änderungen in der tatsächlichen Durchführung führt: Alle im vorangegangenen Schritt definierten und ausgewählten Faktorausprägungen werden in einer Matrix zusammengefasst, wobei jede Faktorausprägung sowohl in einer Zeile als auch einer Spalte steht. Dabei wird gleichzeitig auch eine entsprechende Beeinflussung der Faktoren zueinander definiert, diese kann etwas einfacher als in der originalen Methode ausfallen, da wir durch die Spezialisierung der Faktoren auf

Faktorausprägungen die Werte „Starke negative Beeinflussung“ und „Negative Beeinflussung“ durch ein einfaches „Inkonsistent“ ersetzen können. In den Zellen stehen daher die folgenden Werte:

- -1: Die beiden Faktorausprägungen sind inkonsistent oder konsistente Ausprägungen des gleichen Faktors (der Vollständigkeit halber, diese sollten eigentlich per definitionem inkonsistent sein)
- 0: Zeile und Spalte enthalten die gleiche Faktorausprägung
- 1: Die Faktorausprägung in der Zeile und die in der Spalte sind konsistent, die Relevanz der Faktorausprägung in der Zeile auf die in der Spalte allerdings gering.
- 2: Die Faktorausprägung in der Zeile und die in der Spalte sind konsistent, die Relevanz der Faktorausprägung in der Zeile auf die in der Spalte hoch.

Über die Spalten aggregiert ergibt sich für jede Zeile der Wert der Beeinflussung, definiert durch die Summe aller Einträge größer Null. Dieser fasst zusammen, wie stark eine Faktorausprägung die anderen Faktorausprägungen beeinflusst, zu denen sie konsistent ist. Analog kann dies auch für die Spalten durchgeführt werden, indem eine entsprechende Summierung über die Zahlenwerte erfolgt, diese Summen geben an, wie stark eine Faktorausprägung durch andere beeinflusst wird. Grundsätzlich ist es möglich, dieses Scoring noch deutlich komplexer zu gestalten und den Einfluss aus den Interviews zu erhöhen: Jede Faktorausprägung bekommt dabei einen Basis-Score zugewiesen, der aus den Interviews erwächst. Dieser Score sollte umso höher sein, je öfter eine Faktorausprägung genannt wurde. Zusätzlich kann die Expertise der jeweiligen Expert*innen in der entsprechenden der beantworteten Frage ebenfalls einfaktoriert werden. Wir summieren also im einfachsten Fall auf, wie oft in den Interviews eine Faktorausprägung genannt wurde und können dies im komplexeren Fall mit einem für jede(n) Expert*in festgelegten Wissens-Score für die jeweilige Frage gewichten. Es ist allerdings darauf hinzuweisen, dass die Gefahr besteht sich in vorgetäuschter Genauigkeit zu verlieren. Da schon die Abschätzung der Relevanz in qualitativen Interviews sehr große Unschärfen besitzt, gleichzeitig auch noch Faktoren aus der Literatur hinzugefügt werden, deren Gewichtung ebenfalls schwierig einzuschätzen ist, und speziell in Interview-Situationen natürlich auch der Zufall eine gewisse Rolle spielt, welche Faktoren den Interviewten in der Situation auch tatsächlich einfallen und welche wie prominent präsentiert werden, muss der resultierende Wert mit Vorsicht genossen werden.

Bildung und Auswahl alternativer konsistenter Annahmekombinationen: Basierend auf der vorangegangenen Matrix ist es nun einfach möglich zu erkennen, welche Faktorausprägungen welche anderen beeinflussen, sowie, welche Faktorausprägungen zueinander inkonsistent sind. Dabei ist es natürlich sinnvoll, Szenarien mit einem sehr hohen Score zuerst auszuwählen, allerdings sehen wir es als wenig sinnvoll an, die bspw. Top 5 Szenarien nach Score auszuwählen: Da das Scoring durch starke Konsistenz hochgetrieben wird, kann es weiterhin zu dominierenden Faktorausprägungen kommen, d.h. es besteht die Gefahr der Auswahl sehr ähnlicher Szenarien. Im originalen Ansatz wurde das umgangen, indem auch immer das duale Szenario ausgewählt wird, dies empfinden wir jedoch, wie oben besprochen, als problematisch. Wir schlagen daher vor, hier den Nutzenden der ESA eine gewisse Freiheit einzuräumen und die weiteren Szenarien frei auszuwählen.

Entwicklung und Interpretation der ausgewählten Szenarien: Im Gegensatz zur originalen Methode, können die in der A3-ESA entwickelten Szenarien aus sehr vielen Faktorausprägungen bestehen. Dies erfordert bei der Interpretation und Entwicklung einiges an Fingerspitzengefühl, damit die Szenarien nicht zu überladen wirken. Dieser Schritt verschwimmt oftmals mit dem

nächsten, allerdings geht es hierbei darum, nicht den vollen Kanon an konsistenten Faktorausprägungen eines Szenarios auszuwählen, sondern eine sinnvolle Auswahl zu treffen.

Ausarbeitung der Szenarien und Ableitung von Konsequenzen: Wie auch im originalen Ansatz werden die Szenarien basierend auf der ESA ausgearbeitet und auch ausgeschmückt. Dies ist immer auch mit Unschärfen versehen, es muss auch an dieser Stelle noch einmal nachdrücklich darauf eingegangen werden, dass die ESA keine exakte Wissenschaft darstellt, die die wahrscheinlichsten Szenarien entwickelt, sondern ein Werkzeug dass eine Leitlinie zu möglichen und mehr oder weniger Wahrscheinlichen Entwicklungen liefert. Die Ausgestaltung kann daher noch einmal ein wenig von den Faktorausprägungen abweichen, allerdings ohne zu diesen inkonsistent zu werden. Dabei können bspw. gewisse Faktorausprägungen weggelassen werden, um das Szenario einfacher oder auch freier zu gestalten, bzw. nicht den Fokus zu verlieren. Auch können gleiche Faktorausprägungen in unterschiedlichen Szenarien unterschiedlich interpretiert werden, speziell in Zusammenhang mit anderen Faktorausprägungen, die sich gegenseitig beeinflussen.

Analyse von Herausforderungen, von Effekten auf den Untersuchungsgegenstand und von neuen Anwendungen: Aus den Szenarien können durch Analyse durch Expert*innen die neuen Herausforderungen extrahiert werden, was speziell im Security-Umfeld ein wesentlicher Aspekt ist. Damit kann auch eruiert werden, auf welche künftigen Security-Problematiken besonderes Augenmerk gelegt werden soll. Im Gegensatz zur originalen Methode nach Geschka basieren unsere Szenarien wesentlich auf den Aussagen von Expert*innen im Rahmen von Interviews, in denen oftmals weit mehr als die reinen Faktoren und Faktorausprägungen angegeben wurden, sondern schon von diesen (halbe) Szenarien entworfen wurden. Daher hatten wir, wie bereits in Schritt zwei erwähnt, die Fragestellungen erweitert und explizit nach Problemen, potenziellen Gefahren und Herausforderungen gefragt. Diese werden ebenso in die Analyse inkludiert, so sie mit den Faktorausprägungen des Szenarios konsistent sind. Auch in Bezug auf neue Anwendungen können wir auf Aussagen aus den Interviews zurückgreifen, diese neuen Anwendungen müssen in Hinblick auf die Szenarien nicht nur auf Konsistenz geprüft werden, es muss auch davon ausgegangen werden, dass neue Anwendungen auch zu neuen Gaps führen, die es zu analysieren gilt.

Explorative Szenarioanalyse – Durchführung & Ergebnisse

In diesem Kapitel fassen wir die Ergebnisse der tatsächlich durchgeführten Exploration zusammen, wobei gewisse Learnings in Zusammenhang mit dem Kapitel zu den methodischen Aspekten zu sehen sind, d.h. einige konkrete Ergebnisse der Analyse führten zu methodischen Änderungen und sind daher in beiden Teilen in jeweils unterschiedlicher Beleuchtung dieser Einzelaspekte dargestellt.

Ablauf der Durchführung der ESA

Die explorative Szenarienanalyse wurde im Rahmen des Projekts zweimal, statt wie geplant einmal, durchgeführt:

1. ***Testlauf mit einem ersten Frage-Set zur Interviewdurchführung***: Stark reduzierte Auswahl von zwei Expert*innen. Diese wurden in der zweiten Ausführung nicht mehr befragt, da die Verzerrung durch Wiederholung ähnlicher und teilweise gleicher Fragen sowie durch die anschließende Diskussion zu methodischen Aspekten, nicht in die tatsächliche Exploration übernommen werden sollte.
2. ***Tatsächliche Explorative Szenarienanalyse***: Interview von 12 Expert*innen in einem Remote-Setting. Dadurch konnte eine bessere Vergleichbarkeit der Ergebnisse erzielt werden, als wenn eine Untermenge der Interviews in einem persönlichen Kontext stattgefunden hätte.

Basierend auf den Interviews, wurden, wie im Kapitel zu den methodischen Aspekten ausgeführt, die entsprechenden Faktoren und Faktorausprägungen generiert. Diese wurden in Hinblick auf ihre Konsistenz geprüft, wobei wir hier rein auf Ebene der Faktorausprägungen agierten, da Konsistenzen auf Faktorebene, wie oben dargestellt, oft gar nicht sinnvoll erfassbar sind. Zusätzlich wurden zu den Faktorausprägungen aus den Interviews noch weitere, aus der internen Analyse und der Recherche stammende, Faktorausprägungen hinzugefügt. Zusätzlich wurden noch der folgende Faktor hinzugefügt, der aus einem KIRAS-Vorprojekt stammten und komplett von den aus den Interviews stammenden Faktoren isoliert existiert, d.h. der die Faktoren und Faktorausprägungen aus den Interviews nicht (zu sehr) beeinflusste:

- Sharing Economy – Die Gesellschaft entwickelt sich in eine Richtung, in der viel mehr geshared wird. Dies betrifft nicht nur Daten und die Nutzung von Software oder Unterhaltungsmedien wie Musik oder Filme in der Form von zeitlich begrenzten Lizenzen, sondern speziell auch Gadgets, bis hin zu Autos. Dabei unterscheidet sich Sharing auch von reinen Abo-Services. Sharing kann natürlich mit finanzieller Kompensation einhergehen, kann aber auch auf freiwilliger Basis und ohne Kompensation erfolgen. Wie dies genau aussieht ist nur zweitrangig, wichtig für die Bereiche der AI und der IT-Sicherheit ist, dass potenziell viele Nutzer*Innen gemeinsam Geräte nutzen, damit aber auch AI-Komponenten auf den Daten und den Verhaltensmustern vieler unterschiedlicher Personen trainieren und gleichzeitig ein Nachweis von bösartigem Verhalten Einzelner immer komplexer wird.

Die Faktorausprägungen wurden auf Basis ihrer Relevanz bewertet, wobei das Scoring im Gegensatz zu den ursprünglichen Überlegungen stark vereinfacht wurde und grundsätzlich auf Basis der Nennungen durch die Expert*innen durch einfaches Abzählen erstellt wurde. Dieser extrem einfache Ansatz wurde gewählt, da er (i) sehr einfach durchführbar ist, (i) sehr gut nachvollzogen werden kann, (iii) sich die Ergebnisse nicht grundlegend von Scorings unterschieden, die wesentlich

komplexer die Expertise der einzelnen Expert*innen einfließen ließen, was jedoch im Grunde sehr schwer objektivierbar ist. Eine reine Gewichtung nach Fachbereich schien uns zu stark verfälschend, da auch Sozialwissenschaftler*innen in diesen Grenzbereichen zwischen Technik und Sozialwissenschaften u.U. sehr viel Erfahrung in der Technik besitzen und andererseits speziell Security-Expert*innen sehr starke Expertise in den sozialen Auswirkungen von AI und Security zeigten, die eine Gewichtung rein nach formaler Ausbildung zu einer Scheinobjektivierung werden ließen.

Basierend aus den Faktorausprägungen wurden konsistente Mengen daraus als Szenarien ausgewählt. Dabei wurde ebenfalls zuerst ein recht komplexes Scoring entworfen, das die gewichtete Summe der Expert*innen-Scores in Abwägung mit der inneren Konsistenz als Gewichtung heranzog. Grundsätzlich wurde dies auch für Auswahl des ersten Szenarios als Basisszenario so durchgeführt, allerdings erzeugte dieser Ansatz zu viele zu ähnliche Szenarien. Dies kann auf der einen Seite als sehr positiver Aspekt gewertet werden, zeigt es doch, dass die ESA konsistente Szenarien liefert und generelle Trends und Meinungen gut abbilden und daraus extrapolieren kann. Allerdings ist dieser Vorteil nur dann interessant, wenn man die Menge aller konsistenten Szenarien betrachten will. In der ESA als tatsächliches Mittel der Planung geht es jedoch nicht darum, das Mengengerüst aller konsistenter Szenarien zu analysieren, der Zweck liegt ja darin, einige wahrscheinliche, aber grundsätzlich verschiedene Szenarien zu entwickeln, um daraus dann generelle Gaps abzuleiten und Vorbereitungen für verschiedene Zukünfte treffen zu können. Dies ist ein grundverschiedenes Ziel und um diesem zu genügen, wurden aus den konsistenten Szenarien im Endeffekt bewusst manuell verschiedenen ausgewählt und nicht die 5 mit den höchsten Scores. Zusätzlich ist diese Art von Scores grundsätzlich irreführend und es besteht die Gefahr der sog. „vorgetäuschten Genauigkeit“, d.h. diese Scorings suggerieren eine Genauigkeit in der Arbeit, die durch die hochgradig ungenaue Befüllung der ursprünglichen Daten gar nicht objektiv erreichbar ist.

Im Rahmen dieser Szenarien wurden generelle Gefahren und neue Use-Cases abgefragt. Zusätzlich wurden die Szenarien mit den Use-Cases verschnitten, d.h. es wurde geprüft, wie sich die Use-Cases im Lichte der Szenarien verhalten und welche neuen Gefahren durch die Szenarien bedingt in den Use-Cases auftauchen können.

Zusätzlich wurden im Rahmen des Interviews nicht nur reine ESA-Faktoren und Faktorausprägungen abgefragt, da wir aus der ersten Iteration wussten, dass eine isolierte Abfrage dieser Aspekte ohne Beeinflussung durch das „Szenario im Kopf“ der/des Expert*In nicht realistisch möglich ist. Daher haben wir bewusst auch noch weitere Aspekte im Bereich der Anwendungen und Gefahren abgefragt, die direkt einfließen konnten.

Interview-Phase

Die Interview-Phase diente dazu, möglichst neutral die einzelnen Faktoren und Faktorausprägungen zu eruieren und weitere Wichtige Informationen zur künftigen Entwicklung in die Szenarioentwicklung einbeziehen zu können. Dabei wurden die folgenden vier Dimensionen betrachtet:

- Dimension 1: Aufgaben und Trends in LEAs und der Cyber-Sicherheit
- Dimension 2: Neue Technologien im Bereich KI
- Dimension 3: Rechtliche Entwicklungen im Bereich KI
- Dimension 4: Gesellschaftliche Trends

Dementsprechend wurden auch die Expert*innen ausgewählt, wobei wir auf Diversität in den Geschlechterrollen, sowie in der Altersgruppe achtete, soweit dies möglich und/oder sinnvoll war.

Auswahl der Expert*innen

Die Auswahl der Expert*innen nach den Dimensionen erfolgte nach den folgenden Gesichtspunkten:

- Auswahl anhand der vier Dimensionen, d.h. nicht nur reine Reduktion auf Technik und Security, sondern auch Einbeziehung von Recht und Sozialwissenschaften, speziell Technologiefolgenabschätzung.
- Fokus auf nationale Expertise im Bereich der Security und der Rechtswissenschaften, um die Nutzbarkeit für den Bedarfsträger möglichst hochzuhalten.
- Einbeziehung von international renommierten Expert*innen, auch aus dem Ausland, zur Inklusion der internationalen Dimension.

Insgesamt wurden 20 Expert*innen angesprochen, von denen sich 14 (2 in der Testrunde und 12 in der tatsächlichen ESA) bereit erklärten, die Interviews durchzuführen. Allerdings mussten wir den Teilnehmenden Anonymität zusichern. Dies wurde speziell im Laufe des Jänners und Februars 2025 ein Thema, da hier Expert*Innen explizit nachfragten, dass ihre persönlichen Daten anonym blieben. Die hauptsächliche Begründung war diese, dass in diesen Umbruchzeiten jede Form der Exploration sehr kompliziert ist, Fragen, die über die letzten Jahre relativ einfach zu beantworten waren, auf einmal mit groben Unsicherheiten behaftet sind und unter Umständen wesentliche politische Richtungswechsel bevorstünden, die eine vor einem Monat abgegebene Meinung komplett ins Gegenteil verkehren und den/die Expert*n als „dumm“ dastehen lassen könnte. Daher wollten einige Teilnehmenden nicht mehr namentlich genannt werden, da Aussagen aus Ende 2024 Mitte 2025 schon ganz anderen Kontext besitzen könnten und speziell Gefahren und Chancen sich durch die politischen Turbulenzen massiv verändern können.

Wir haben daher kein*e der Expert*Innen in dieser Studie namentlich genannt, die Namen liegen jedoch intern vor und wurden auch dem Bedarfsträger im Sinne der Nachvollziehbarkeit und Kontrolle der tatsächlichen Expertise genannt. Auch werden die originalen Fragebögen nicht in der Studie angehängt, sie stehen aber dem Bedarfsträger zur Einsicht zur Verfügung.

Leider kam diese zusätzliche Komplexität erst im Nachgang der Interviews zu Tage, d.h. die Interviews wurden großteils vor den derzeitigen politischen Turbulenzen durchgeführt, daher konnten wir dies nicht vorher bereits abfangen. Allerdings sind wir uns sicher, dass es derzeit sehr schwierig ist, und speziell im Februar noch viel schwieriger war, Expert*innen zu definitiven Aussagen bezüglich künftiger Entwicklungen in einem derart heiklen Umfeld zu bewegen.

Außerdem führt die derzeitige politische Umwälzung natürlich dazu, dass viele Aspekte der Entwicklung, wie bspw. bezüglich starker Regulierung in der Europäischen Union, sich sehr stark von den antizipierten Einflüssen unterscheiden werden können. Dies möchten wir damit begründen, dass auch unsere Expert*Innen diese politischen Schwierigkeiten nicht realistisch vorhersehen konnten und daher die Ergebnisse weiterhin ihre langfristige Gültigkeit besitzen, aber im Lichte der jüngsten Entwicklung teilweise deutlich adaptiert werden müssen. Manchmal beinahe täglich.

Die zwölf Interviewpartner verteilten sich folgendermaßen auf die vier Dimensionen:

- D1 – LEAS und Security: 5
- D2 – AI: 4
- D3 – Rechtswissenschaften: 1
- D4 – Sozialwissenschaften: 2

Die Ungleichgewichtung rührt zum einen daher, dass manche Expert*Innen in zwei der Gebiete Expertise besitzen (speziell Security & AI, Security & Sozialwissenschaften, sowie AI und Sozialwissenschaften) und wir immer nur die Hauptexpertise als Auswahlkriterium heranzogen, zum anderen aufgrund des unterschiedlichen Antwortverhaltens, so waren Security-Expert*Innen responsiver.

Fragenkatalog und Ablauf der Interviews

Bereits in der ersten Durchführung fiel auf, dass es in einer Interview-Situation nicht möglich ist, die Eruierung der Faktoren losgelöst von ihren Faktorausprägungen und von den Präferenzen der Faktorausprägungen der Expert*innen zu trennen, d.h. in einer echten Interview-Situation, in der auch ein Gespräch anhand von Fragen durchgeführt wird, und nicht rein passiv eine Kreuzerlliste an vorausgewählten Faktoren abgearbeitet wird, schwingen die von den Expert*innen antizipierten Szenarien nicht nur unterschwellig mitschwingen, sie kommen sehr stark zu Tage und dominieren das Interview. Daher ist die Frage nach Faktoren, wie im vorherigen Kapitel bereits detailliert dargestellt, relativ sinnlos, die Expert*innen liefern nicht nur konkrete Faktorausprägungen, sie liefern zumindestens in manchen Aspekten komplette Teilszenarien. Dabei wurde allerdings beobachtet, dass diese nicht immer komplett in sich geschlossen und widerspruchsfrei sind, d.h. Teilszenarien, die in der Beantwortung einer Frage auftauchten und dort vehement dargelegt wurden, sind nicht immer kompatibel zu Teilszenarien gewesen, die im Rahmen der Beantwortung anderer Fragen mitgeteilt wurden.

Dies wiederum spricht sehr stark dafür, den gewählten methodischen Ansatz auch wirklich konsequent bis zum Ende durchzuführen, indem aus den Antworten Faktorausprägungen isoliert und diese abstrakt auf Konsistenz hin geprüft werden. Dadurch können Szenarien entwickelt werden, die wesentlich besser und vor allem konsistenter als die Meinung der Expert*innen sind.

Ein weiterer überraschender Aspekt aus Ergebnissicht war der, dass wir erwartet hatten, dass sich Expert*Innen aus den Sozialwissenschaften wesentlich kritischer über AI und speziell der Anwendung von AI durch LEAs äußern würden als dies Techniker*Innen, und speziell AI-Expert*Innen, tun würden. Diese Annahme konnte nicht bestätigt werden, zu unserer Überraschung waren gerade Expert*Innen aus dem Feld der AI kritisch gegenüber der unkritischen Anwendung von AI, nicht nur im Bereich des Law-Enforcements, sondern auch in anderen technischen Anwendungen. Der Tenor war sehr stark ausgeprägt, dass Menschen Maschinen nur dann kontrollieren und letztinstanzliche Entscheidungen treffen können, wenn diese Menschen auch genügend Erfahrung in der Durchführung der jeweiligen Aufgaben hätten – die großflächige Automatisierung von einfachen Tätigkeiten und das damit stattfindende „Unlearning“ könnte ein Problem zeitigen, indem es die Ausbildung zu senioren Tätigkeiten, d.h. der nicht durch AI automatisierbaren Expert*Innen, verhindere.

Auch der Fragenkatalog wurde zweimal überarbeitet, einmal intern im Rahmen eines Workshops zusammen mit dem Bedarfsträger, ein anderes Mal direkt im Nachgang der ersten ESA-Runde mit zwei ausgewählten Expert*Innen.

Ursprünglich waren maximal 10 Fragen geplant, mit freien Antworten und einem Fokus auf kurze Antwortmöglichkeiten. Der Katalog sollte vorher zugeschickt werden. Die Expert*Innenmeinung

sollte mit Gewichten (1,3,5 nach Maßgabe der Expertise und der Expertise im Fachgebiet) versehen werden. Der ursprüngliche, im Rahmen des Workshops überarbeitete, Fragenkatalog hatte die folgende Gestalt:

- Welche neuen Aufgaben sehen Sie im Bereich des Law-Enforcements durch AI auf LEAs zukommen?
- Wie wird AI LEAs bei ihren Aufgaben unterstützen können?
- Welches Gefahrenpotenzial sehen Sie im Einsatz von AI für die Bevölkerung?
- Wie wird sich die Regulierung von Software im Allgemeinen und AI im Speziellen in der EU weiterentwickeln?
- Wie sehen Sie die Akzeptanz von AI in der Gesellschaft in der Zukunft?
- Welche neuen KI-gestützten Anwendungen mit hohem Zukunftspotenzial erwarten Sie sich?
- Welche gesellschaftlichen Trends in Bezug auf IT sehen Sie in der näheren Zukunft?
- Wie wird AI Ihrer Meinung nach das tägliche Leben von Menschen beeinflussen?

Aus der Testrunde der ESA folgten die folgenden Learnings, die direkt in die Gestaltung des finalen Interview-Fragebogens einflossen:

- ***Learning 1 - Interviewfragen zu breit***: Die Fragen waren viel zu unkonkret, teilweise luden sie dazu ein, gleich in der ersten Frage die gesamte Meinung zur zukünftigen Entwicklung von AI, LEA und der Gesellschaft als Ganzes als gebündeltes Szenario weiterzugeben, was (i) nicht Sinn des Interviews war und (ii) die Extraktion von Faktoren sehr erschwerte und (iii) auch den Interviewenden, der ja auf Basis der gewonnen Informationen erst die Szenarien neutral entwickeln sollte, in eine gewisse thematische Richtung trieb. Grundsätzlich wäre es wünschenswert, wenn Interviewdurchführung und Auswertung von verschiedenen Personen durchgeführt werden, allerdings stellt dies die nachgelagerte Bearbeitung wiederum vor Herausforderungen, wie später in diesem Kapitel beschrieben.
- ***Learning 2 - Interviewfragen zu fokussiert***: Teile der Fragen waren zu fokussiert auf LEA, hier wurde der Fragenpool etwas erweitert, um auch andere gesellschaftliche Faktoren besser einzubeziehen.
- ***Learning 3 - Ein großer thematischer Fokus***: Grundsätzlich war der gesamte Themenpool sehr groß gewählt, von direkten und konkreten Auswirkungen auf LEAs, bis hin zu Fragen der Regulierung von AI und Technik, aber auch der gesellschaftlichen Akzeptanz. Dies steht auch in einem gewissen Gegensatz zur Methode von Geschka, die sich grundsätzlich auf eine einzige Frage fokussiert und diese in Bezug auf Faktoren untersucht. Allerdings sind alle diese Teilfragen für den Bedarfsträger von hoher Relevanz gewesen, eine wiederholte Abfrage in verschiedenen Interviews mit den gleichen Expert*Innen war realistisch kaum möglich. Es wäre natürlich möglich gewesen, für jede der inhaltlichen Teilfragen eine eigene ESA durchzuführen und für jede die Faktoren einzeln abzufragen, dies wurde allerdings von uns verworfen, da (i) der Aufwand auf Seite der Interviews rapide steigt, (ii) die Themenkomplexe zu stark verwoben sind, die Antworten der Expert*Innen auf eine Frage daher sehr stark durch die vorangegangenen Diskussionen und Antworten und den

darin entworfenen „Szenarien im Kopf“ beeinflusst gewesen wäre, diese realistisch referenziert und auf sie Bezug genommen hätte, und (iii) bei Nutzung getrennter Mengen an Expert*Innen der benötigte Pool sehr groß geworden wäre und gleichzeitig wiederum an Vergleichbarkeit verloren hätte. Daher haben wir uns entschieden, mit dieser Herausforderung umzugehen und etwaige Unschärfen und gegenseitige Beeinflussungen hinzunehmen. Hierbei hat es such bezahlt gemacht, dass die Interviews direkt geführt wurden und nicht einfach vorausgefüllte Fragebögen verteilt wurden, da der Interviewer entsprechend steuernd eingreifen konnte.

- ***Learning 4 – Szenarien als Antwort***: Wie schon eingangs erwähnt, haben Expert*Innen in einem Gebiet natürlich selbst ein Bild der künftigen Entwicklung im Kopf, oft sogar mehrere Szenarien basierend auf externen Faktoren und eintretenden oder ausbleibenden Rahmenparametern. Es ist sehr schwierig, diese Szenarien auszublenden, wenn in Interviews konkret nach Auswirkungen einer Technologie auf die Zukunft gefragt wird. Wir haben uns damit beholfen, die Interviews umzudrehen und, wie unten in den endgültigen Fragen gut ersichtlich, zuerst sehr abstrakte Faktoren zu erfragen. Dennoch mussten wir feststellen, dass auch bei sehr konkreten Fragen nach Faktoren, der Interviewer stark steuernd eingreifen musste, weil die Expert*Innen nicht Faktoren, sondern gleich Faktorausprägungen lieferten Diese wurden auch gleich begründet in ihrer Auswahl und warum sie für wichtig erachtet wurden, was wiederum darin mündete, dass sehr weitschweifig im Interview Szenarien der Zukunft direkt von den Expert*Innen entworfen wurden. Dies konnte durch eine disziplinierte Form der Interviewführung in geregelte Bahnen geleitet werden, wobei sich wiederum die Durchführung eines ersten Testlaufes bezahlt gemacht hatte. Zusätzlich ergaben die Umsortierung und direkte Abfrage der Faktoren typischerweise sehr gute erste Antworten, die gut in der ESA genutzt werden konnten.

- ***Learning 5 - Interviews auf Basis einer ersten Runde an Einflussfaktoren***: In der originalen Methodik nach Geschka werden die Faktoren nicht notwendigerweise aus den Interviews erhoben, sondern bspw. Im Rahmen einer Literaturrecherche. Dies dient dazu, dass offensichtlich fehlende Expertise oder vergessene Faktoren nicht ignoriert erden. Wir haben dies entsprechend im Nachgang an die Interviews inkludiert. Allerdings haben wir bewusst auf Interviews basierend auf einer Vorauswahl der Faktoren verzichtet: Dies hätte unserer Meinung nach einen starken Bias in die Interviews eingebracht. Die grundlegende Idee der ESA war, das Wissen der Expert*Innen für die Szenarienentwicklung zu gewinnen. Eine Vorauswahl hätte aber die Expert*Innen zu reinen Gutachter*Innen über rein im Projektteam durchgeführte Arbeiten gemacht, Je nach Temperament der Expert*In hätte diese/r vielleicht noch weitere Faktoren einbezogen, die freie Interviewtätigkeit mit der Breite der Antwortmöglichkeiten zeitigte allerdings auf jeden Fall ein unvoreingenommeneres Bild der Meinung der Expert*Innen als dies jede Vorauswahl ermöglicht hätte. Wir haben daher bewusst diese Abweichung in der Methodik in Kauf genommen und die Methodik auch entsprechend adaptiert.

- ***Learning 6 - Positive Fragen nach Chancen ergeben (neue) Use-Cases***: Die ursprünglichen Fragen waren aus der Sicht von Security-Expert*Innen gestellt und fokussierten sich sehr stark auf die Gefahren von AI für die Gesellschaft, aber auch als Herausforderung für Law Enforcement. Dabei wurden die positiven Aspekte von AI auf die Arbeit von LEAs kaum konkret abgefragt, was zu einem sehr stark negativen Bias führte, der auch von den Interview-Partnern bemerkt wurde, da diese am Ende bewusst auf die positiven Aspekte von AI auf LEAs zu sprechen kamen und diese als sehr hoch bewerteten. Speziell wurden auch neue Use-Cases aus den Antworten extrahiert, was in der ursprünglichen Methodik

nicht vorgesehen war, da wurde lediglich die Auswirkung der Szenarien auf die Use-Cases untersucht.

- ***Learning 7 - Umsetzung der Antworten recht aufwändig***: Der Aufwand der Extraktion von Faktoren und Faktorausprägungen aus den freien Antworten ist nicht zu unterschätzen, wobei auch Vieldeutigkeiten ein Problem darstellen können. Dabei ist die Durchführung von Interviews durch einen Experten als Interviewenden von unschätzbarem Vorteil, speziell, wenn dieser auch die nachfolgende Auswertung durchführt, da so direkt im Interview Unklarheiten beseitigt und Nachfragen gestellt werden können.
- ***Learning 8 – Keine Abfrage nach Faktoren und Faktorausprägungen***: In der ersten Iteration wurde rein indirekt gefragt, um aus den Antworten zu freien Themen auf die Faktoren schließen zu können. Dies wiederum hat den Raum für Interpretation durch den Interviewenden stark erhöht und führte zu einer minderen Qualität in der Extraktion der Faktorausprägungen. Speziell bei unklaren Faktoren wurde das Feld der möglichen Faktorausprägungen sehr breit, was zu vielen Möglichkeiten von geringer Relevanz und geringer Basis in der Expert*Innenmeinung führte. Wir haben daher entsprechende Fragen hinzugefügt und an den Anfang gestellt.

Folgend die finale Version der Interviewfragen:

1. Was sind wichtige Faktoren für die weitere Entwicklung von KI?
2. Was sind wichtige Faktoren für die gesellschaftliche Akzeptanz der KI-Nutzung?
3. Was sind wichtige Faktoren für die polizeiliche Nutzung von KI?
4. Welches Gefahrenpotenzial sehen Sie im Einsatz von AI für die Bevölkerung?
5. Wie sehen Sie die Akzeptanz von AI in der Gesellschaft in der Zukunft?
6. Welche gesellschaftlichen Trends in Bezug auf AI sehen Sie in der näheren Zukunft?
7. Wie wird AI Ihrer Meinung nach das tägliche Leben von Menschen beeinflussen?
8. Welche neuen Herausforderungen durch die Nutzung von AI sehen Sie auf LEA s zukommen?
9. Wie wird AI LEAs bei ihren Aufgaben als Werkzeug unterstützen können?
10. Wie wird sich die Regulierung von Software im Allgemeinen und AI im Speziellen in der EU weiterentwickeln?

Die durchschnittliche Interviewdauer betrug rund 30 Minuten für den inhaltlichen Teil, oft noch abgerundet mit weiteren 30 Minuten allgemeine Gespräche über die Zukunft und AI in der Zukunft im Speziellen.

Auswertungsphase

Basierend auf den Interviews wurden die Fragebögen direkt durch den Interviewenden ausgewertet. Wir entschieden uns dabei, dass dies geschlossen im Nachgang des letzten Interviews

durchgeführt werden sollte, da so eine Konsistenz in der Auswahl erreicht wurde, die über Wochen hinweg kaum zu halten gewesen wäre.

Isolierte Faktoren und Faktorausprägungen

Die Faktoren und Faktorausprägungen wurden nicht nur aus den Fragen isoliert, die diese direkt abfragten, sondern es wurden speziell auch indirekt genannte Faktoren und deren Ausprägungen aus den anderen freien Antworten extrahiert, bspw. wenn Entwicklungsmöglichkeiten und Parameter als Begründung für Restriktionen durch Regularien angeführt wurden. Dies bedeutet natürlich eine gewisse Ungenauigkeit in der Aggregation, da manchmal sehr ähnliche Antworten gegeben wurden, die sich nur in Formulierung, oder aber geringen Teilaspekten unterschieden. Gleichzeitig mussten gegeben Faktoren teilweise auch aufgeteilt werden, bspw. wurde das Thema der Transparenz von vielen Expert*Innen genannt, allerdings mit unterschiedlicher Bedeutung, je nach Hintergrund. Teilweise benutzten Expert*Innen auch das gleiche Wort in unterschiedlichen Bedeutungen in Antworten zu unterschiedlichen Fragen. Diese natürliche Unschärfe ist inhärent und kann nur durch sorgfältiges Nachfragen und einen kundigen Interviewer eingegrenzt werden, es muss aber herausgestellt werden, dass auch diese Person natürlich bewusste und unbewusste Entscheidungen trifft, welche Antworten wie aggregiert und geteilt werden und damit die resultierenden Faktorausprägungen beeinflusst.

Aufschlüsselung der prominentesten Antworten nach Fragen

Im folgenden Abschnitt geben wir eine Analyse der prominentesten Antworten pro Frage wieder, so wie sie direkt aus den Interviews extrahiert wurden.

Frage 1 – Was sind wichtige Faktoren für die weitere Entwicklung von KI?

- Top-Faktoren
  - Intensität der Regulierung und Strenge der Auslegung von Regularien
  - Ressourcen(verbrauch) & Kosten von AI
  - Weitere Erfolge aber auch Limitierungen in Algorithmik und bezüglich Optimierung
  - Qualität der Ergebnisse der Systeme
- Zusätzliche Aspekte
  - Die Datenverfügbarkeit war überraschenderweise ein wenig prominent genanntes Thema, allerdings kann dies auch in der Freiheit der Fragebeantwortung begründet sein. Hierzu nehmen wir an, dass ein Multiple-Choice-Test mit Antwortmöglichkeiten/Scorings ein anderes Ergebnis gebracht hätte, wir sind daher auch mit dieser Art der Datenerhebung sehr zufrieden, weil sie wiedergibt, was den Expert*Innen selbst zu dem Thema einfiel.
  - In spätere Interviews wurde Geopolitik und vor allem das Thema der geopolitischen Stabilität, bzw. die Rolle und Anhängigkeiten von Europa viel mehr Thema, was in gewisser Weise eine Verzerrung darstellt.

Frage 2 – Was sind wichtige Faktoren für die gesellschaftliche Akzeptanz der KI-Nutzung?

- Top-Faktoren
  - Nachvollziehbare oder kontrollierbare Ergebnisse
  - Gezielter & transparenter Einsatz
  - Missbrauch von AI du das Thema „Fake News“
  - Tatsächlicher Nutzen nach dem Abklingen der ersten Euphorie

Frage 3 – Was sind wichtige Faktoren für die polizeiliche Nutzung von KI?

- Top-Faktoren
    - Zuverlässige Datenquellen
    - Sicherheit & Robustheit der AI-Anwendungen
    - Transparenz in der Verarbeitung
    - Reproduzierbarkeit & Kontrolle
    - Rechtliche Absicherung
- Viele unterschiedliche weitere Faktoren mit geringer Zahl an Nennungen
    - Thematik des Missbrauchs von AI durch LEAs
    - Tatsächlicher Nutzen statt Selbstbeschäftigung
    - Gibt es einen „Human in the loop“, bzw. Menschen als Letztentscheider*Innen
    - Spätere Interviews fielen durch mehr Geopolitik auf.

Frage 4 – Welches Gefahrenpotenzial sehen Sie im Einsatz von AI für die Bevölkerung?

Hier kam es zu extrem unterschiedlichen Antwortmustern, von sehr positiv bis stark negativ gefärbt, von sehr abstrakten Gefahren bis hin zu sehr konkreten Beispielen. Wesentliche Beispiele werden auch weiter unten diskutiert, speziell den Top-Picks haben wir im Laufe dieses Kapitels noch genauere Darstellungen und Analysen gewidmet.

- Top-Gefahren
    - Unlearning & Overreliance, Entmündigung
    - Verstärkung von Biases und Fehler
    - Übermäßiges Vertrauen in AI
- Weitere Beispiele
    - Near-KiPo: Die Erstellung von künstlichem pornographischem Material, bei dem es sehr unklar ist, ob es bereits den Tatbestand der Kinderpornographie erfüllt, oder gerade noch im legalen Bereich ist.
    - Einzelschicksal vs. Optimierung: Wie sehr wird die Beachtung von Einzelschicksalen dem Gedanken der Optimierung unterworfen. Wie sehr wird die Optimierung zum Selbstzweck, ohne zu überprüfen, ob der Zweck der optimierten Abläufe überhaupt sinnvoll ist.

Frage 5 – Wie sehen Sie die Akzeptanz von AI in der Gesellschaft in der Zukunft?

- Prominenteste Faktorausprägungen
    - Hoch
    - Unvermeidbar
- Allerdings oftmals als abhängig von anderen Faktoren genannt:
    - Qualität der AI-Systeme und Ergebnisse
    - Anwendungsbereich von AI
    - Geopolitik und die Frage, wer die Kontrolle über die Systeme hat
    - Eventuell auftretende Skandale, die den Ruf von AI beschädigen
- Speziell wurde auch das Thema genannt, dass die KI als Sündenbock für menschliche strategische Fehlentscheidungen herhalten könnte.

Frage 6 – Welche gesellschaftlichen Trends in Bezug auf AI sehen Sie in der näheren Zukunft?

- Top-Trends

    - Auswirkung auf Berufe und den Arbeitsmarkt
    - Deep Fake und Social Media-Integration
    - AI wird unsichtbar & allgegenwärtig
    - Bildung in dem Bereich steigt
- Weiteres unerwartete Trends
    - Sündenbock KI (siehe Frage 5)
    - Einfluss auf menschliches Beziehungsverhalten (siehe später im Kapitel zu ausgewählten Anwendungen und Herausforderungen)

Frage 7 – Wie wird AI Ihrer Meinung nach das tägliche Leben von Menschen beeinflussen?

- Top-Meinungen
    - Kompetenzverlust bezüglich Wissens und Entscheidungsfähigkeit (siehe später im Kapitel zu ausgewählten Anwendungen und Herausforderungen)
    - Art des Umgangs mit Wissen
    - Existenz von Intelligente Assistenten und Vernetzung
    - Omnipräsenz der AI
    - Extreme Ausrichtung hin zu Effizienz in vielen Lebensbereichen

Frage 8 – Welche neuen Herausforderungen durch die Nutzung von AI sehen Sie auf LEA s zukommen?

- Top-Herausforderungen
    - Alte Angriffe mit besseren (AI-basierten) Methoden wie Spam und Phishing
    - Cyber-Kriminelle haben Vorsprung, da sie sich nicht an Regularien und Prozesse halten müssen
    - Eintrittsschwelle zu Cyber-Crime deutlich niedriger, da viele einfach anzuwendende Tools vorhanden sind
- Weitere interessante Herausforderungen
    - AI Noise – Kriminelle generieren große Mengen an Noise-Daten durch AI, die realistisch und/oder interessant aussehen, um durch die Flutung von Systemen mit solchen Daten Forensik und Detection-Mechanismen der LEAs auszuhebeln, bzw. zu überfordern.
    - Pattern Matching auf Polizeidaten, um die Arbeitsweise und die Ressourcenverortung er LEAs zu analysieren und gezielter kriminell arbeiten zu können.

Frage 9 – Wie wird AI LEAs bei ihren Aufgaben als Werkzeug unterstützen können?

Hier kam es zu vielen verschiedenen Antworten.

- Top-Picks
    - Analysetätigkeiten aller Art, angefangen von Daten aus Social Media, bis hin zu Netzwerktraffic
    - Automatisierung von Tätigkeiten, speziell in der Verwaltung, in Bezug auf das Schreiben von Berichten und organisatorische Aufgaben
    - Umgang mit Wissen, speziell im Wissensmanagement und der Wissensbeschaffung

- Interessante Einzelmeinungen
    - Erstellung von Fahndungsfotos aus Personenbeschreibungen anstelle der derzeit üblichen Vorgangsweise könnte massiv Ressourcen in diesem Bereich sparen und gleichzeitig schneller besserer Ergebnisse zeitigen.
    - Riot-Control im Sinne der Auswertung und Analyse von Personenströmen und Verhaltensmuster von Personen
    - Bodycam-Auswertung und Summary – Automatisierte Sichtung von Body Cams und Zusammenfassung der Filme in Berichten.

Frage 10 – Wie wird sich die Regulierung von Software im Allgemeinen und AI im Speziellen in der EU weiterentwickeln?

- Top-Ansicht: Einpendeln auf einen guten Mittelweg zwischen regulatorischem Schutz von Personen auf der einen Seite und der Ermöglichung von Innovation und Forschung, aber auch neuen Anwendungen auf er anderen. Allerdings waren sich die Expert*Innen, die diese Meinung vertraten, relativ einig dabei, dass dies ein iterativer Prozess werden würde, bei dem das Pendel immer wieder in beide Richtungen ausschlagen wird.
- Allerdings wurde ein breites Spektrum an Antworten gegeben, wobei die Tendenz Richtung weniger Regulierung ging, speziell am Ende der Interviewperiode, als auch geopolitische Fragen zunehmend wichtiger wurden. Dieser Shift in der Geopolitik, der durch das Verhalten der amerikanischen Regierung Ende Jänner bis Mitte Februar offenbar wurde, hat das Antwortverhalten der Expert*Innen durchaus verändert. Gerade aufgrund dieses Shifts wurde argumentiert, dass es sich Europa nicht länger leisten kann, alles zu streng und vorsichtig zu regulieren.
- Weitere wichtige Themen waren IPR und Haftungen, erstere speziell in Hinblick auf die Vergütung der Werke, die von AI-Herstellern für das Training der Systeme genutzt wurden/werden, was bisher von den großen Vertretern eher ignoriert wurde, letzteres speziell in Hinblick auf Fehler oder problematische Konsequenzen aus AI-Nutzung, auch in Hinblick auf Human-in-the-Loop-Systeme, in denen der Mensch eigentlich als letzte Instanz eingebunden ist, diese aber entweder aufgrund des Unlearning-Problems nicht mehr wahrnehmen kann, oder aber die zeitliche Komponente eine menschliche Entscheidung nicht mehr erlaubt.

Ausgewählte Anwendungen und Risiken aus den Interviews

Aus den Interviews konnte auch eine Reihe von neuen Anwendungen, speziell im Law Enforcement, aber auch Risiken extrahiert werden. Im Folgenden wollen wir auf die zwei prominentesten genannten Anwendungsgebiete, sowie zwei wesentliche Quellen für Risiken eingehen.

Automatisierung: Die natürlich am meisten genannte „neue“ Anwendung ist die logische Fortführung der bisherigen Erfolgsgeschichte der IT, die Automatisierung von Tätigkeiten, speziell solchen, denen die Menschen wenig Wert beimessen. Dabei wurden vor allem die folgenden konkreten Anwendungsfälle im polizeilichen Umfeld genannt:

- ***Datenanalyse und Zusammenhänge***: Dies war unmissverständlich die Top-Anwendung, wenig überraschend ist Pattern-Matching und die Datenanalyse ein Thema, dem in der polizeilichen Arbeit viel Potenzial zur Performancesteigerung beigemessen wird. Dies trifft

auch explizit auf genannte Anwendungen im Bereich des Cyber Crimes zu, speziell auch die Auswertung von Sensoren und großen Datenmengen.

- ***Standardtätigkeiten***: Tätigkeiten aus der Verwaltung, wie das Ausformulieren von Berichten, das Erstellen von Übersetzungen und Zusammenfassungen, aber auch weitere formale und organisatorische Tätigkeiten, die neben der eigentlichen Aufgabe im polizeilichen Dienst wahrgenommen werden müssen.

- ***Security & Monitoring***: Klassische Anwendungen wie Überwachungskameras und deren Auswertung, speziell in der Triage die Erkennung und Filterung unrelevanter Informationen, damit sich menschliche Analysten auf die interessanten und/oder uneindeutigen Aspekte fokussieren können.

- ***Social Media***: Analyse von Cliquen und Kontakten in Sozialen Netzwerken, aber auch Inhalte von Social Media Posts, speziell in der Auswertung von Kontakten und Delikten wie Hate-Crimes. Auch sehr gut für die Analyse von Trends geeignet.

- ***Wissensbeschaffung***: Scrappen des Netzes nach Informationen, sowohl generischer, als auch spezifischer Natur. Dabei wurde auf das Problem hingewiesen, dass Angreifer damit arbeiten könnten, hidden content in Websites einzubetten, der für menschliche Besucher nicht erkennbar ist, aber direkt Daten in polizeiliche Überwachungssysteme einbringt. Zweck dieser Übung könnte sein, unverfängliche Websites auf Überwachungslisten zu bringen und damit eigene Aktivitäten zu tarnen, die Besucherunverfänglicher Websites zu kriminalisieren, oder aber einfach die Ressourcen der Polizei auszuschöpfen (DDOS).

- ***Psychologisch belastende Tätigkeiten übernehmen***: Hierbei wurde vor allem die Nutzung zur Sichtung von illegaler Pornographie, speziell solcher mit Kindern, genannt, die für die menschlichen Ermittler*Innen psychisch extrem belastend ist. Natürlich müsste hierbei auch wieder letztinstanzlich ein Mensch entscheiden, aber es wurden einige Anwendungsfälle angegeben, die durch Vorfilterung bereits einiges an Belastung reduzieren könnte, bzw. gewisse Tätigkeiten, wie die quantitative Auswertung maschinell erledigbar machen könnten.

- ***Unteres Management***: Eng verwandt mit der Übernahme von Verwaltungstätigkeiten könnten auch viele konkrete Planungstätigkeiten durch den Einsatz von AI profitieren und die damit beschäftigten Personen für andere Zwecke freispielen.

- ***Rechtliche Analyse / Paragraphen-Matching***: Ein öfter genannter Anwendungsfall ist eng verwandt mit der Berichtserstellung: Auch wenn Polizisten sehr gute Kenntnisse der grundsätzlichen Rechtsmaterie besitzen, so ist in der konkreten Ausformulierung immer auf die konkreten Gesetzesstellen zu verweisen. Dies zu Matchen könnte eine einfach zu erreichende Performance-Steigerung zeitigen, speziell in Hinblick auf selten angetroffenes Recht oder Fragen aus niedrigschwelligen Problembereichen wie Nachbarschaftsstreitigkeiten und dergleichen, die oft sehr spezielle Rechtsfragen aufgreifen. Ein automatisiertes Matching könnte es dem Ermittler erlauben, sehr rasch die geeigneten Gesetzesstellen herauszufinden und einzuarbeiten, Dies würde auch für andere Ebenen des Rechtssystems, bspw. Anwälte, Staatsanwälte und Richter*Innen gelten, die ebenfalls von solchen Systemen profitieren könnten, auch wenn die letztinstanzliche Entscheidung in diesem Fall gar nicht von der AI beeinflusst werden soll, da es hierbei lediglich um eine intelligente Version des Nachschlagens geht. Damit sollte dies auch rechtlich relativ wenig problematisch sein, da keine direkten Effekte auf die Rechtsprechung geplant wären.

Interaktive Assistenz: Ein weiterer spannender Anwendungsfall wäre die Nutzung als interaktive Assistenz. Speziell in vielen kleinen Wachstuben ist es nicht möglich, alle irgendwann notwendigen Kompetenzen vorzuhalten. Speziell wurde dabei das Thema der Bearbeitung von Cyber-Crime in sehr kleinen Wachstuben genannt, wo viel aktuelles Wissen notwendig ist, das auch entsprechend ausgebildetes Personal notwendig macht. Hierzu könnte den Polizeibeamten eine interaktive Assistenz zur Verfügung gestellt werden, die Anleitungen und Hilfestellungen geben kann, was zu tun ist, bzw. wer zu involvieren ist, falls Assistenz und Beamte/r an ihre Grenzen stoßen. Dies könnte als First Level Support für viele seltene Aufgaben herangezogen werden und hätte den Vorteil der einfachen Nutzung durch sprachliche Eingabe mit anschließender Interpretation und Wissensvermittlung mittels eines LLMs. Diese Anwendung wäre sehr niederschwellig in der Entwicklung und könnte rasch einsatzbereit sein, wobei natürlich auch hierbei die betroffenen Beamt*innen dazu ausgebildet werden müssen, Anweisungen nicht direkt und ohne Hinterfragen zu übernehmen.

Im Folgenden wollen wir noch zwei wesentliche genannte Risikogruppen darstellen, die sehr prominent in den Interviews genannt wurden:

Unlearning & Overreliance, Entmündigung: Bei diesem Risiko geht es darum, dass Menschen Tätigkeiten verlernen. Dies ist grundsätzlich kein Problem, sehr viele Kulturtechniken vergangener Zeiten werden durch moderne Menschen nicht mehr beherrscht, bspw. Feuermachen mit Holzbogen oder das Reiten. Es wird allerdings zu einem Problem, wenn Tätigkeiten nicht mehr beherrscht werden, die weiterhin wichtig sind. Zusätzlich erzeugt jeder Verlust einer derartigen Kulturtechnik neue Abhängigkeiten. Im Fall der Automatisierung durch KI ist abschätzbar, dass erhebliche Probleme auftreten könnten, wenn viele Tätigkeiten nicht dennoch weiterhin von Menschen beherrscht werden:

- ***Hand-Over Maschine → Mensch***: Ein wesentlicher Punkt, der speziell auch in Bezug auf die Nutzung Autonomen Fahrens erwächst, ist das Problem des Hand-Overs von Maschine zu Menschen: Sollte sich die Maschine (das automatische Auto) nicht mehr aus einer Situation heraussehen und diese nicht mehr beurteilen können, so soll der Mensch eingreifen und das Problem, quasi als Letztinstanz, lösen. Dabei muss sich die Frage stellen, ob (a) ein Mensch überhaupt in der Lage ist, Entscheidungen zu treffen mit denen eine entsprechend gut trainierte AI, die auf eine Vielzahl an Sensorinformationen quasi instantan zugreifen kann, überfordert ist, sowie (b) ob ein Mensch, der es gewohnt ist eine Tätigkeit nicht auszuführen, diese einfach so instantan übernehmen kann. Dabei spielt vor allem das Thema der Aufmerksamkeitsspanne des Menschen, aber auch das Thema der regelmäßigen Übung von Tätigkeiten eine sehr große Rolle.

- ***Der Mensch als finale Entscheidungsinstanz***: Speziell in Arbeiten zu Trustworthy AI wird oft gefordert, dass der Mensch die letztendlich entscheidenden Instanz zu bleiben hat, speziell in Themen wie medizinischer Diagnostik, aber auch in Autonomous Weapon Systems. Hier muss die Frage gestellt werden, ob Menschen, die solche Entscheidungen schon sehr lange nicht durchgeführt haben, überhaupt in der Lage sind, die Vorschläge einer AI kritisch zu beurteilen, oder ob es zu einem Zustimmungsautomatismus kommt, bei dem Vorschläge der AI im Endeffekt letztinstanzlich abgenickt werden, ohne diese zu hinterfragen. Dabei spielen zwei Probleme eine Rolle: (a) Ob Menschen überhaupt noch die Kenntnis besitzen, die Lage und die Vorschläge entscheiden zu können und (b) ob sie dies auch tun, oder das übermäßige Vertrauen in AI sie daran mental hindert, Vorschläge zu hinterfragen. Wesentlich in zeitkritischen Anwendungen ist auch die Frage der

physischen Möglichkeiten eines Menschen, derart viele Informationen aufzunehmen, wie sie einer AI im Vorschlagswesen quasi instantan zur Verfügung stehen und deren Auswertung in der notwendigen Geschwindigkeit durchzuführen, d.h. on Kontrolle in Entscheidungsfindungen, die in Sekundenbruchteilen geschehen müssen, nicht nur eine Vortäuschung einer menschlichen Kontrolle darstellen.

- ***Kurzfristiger Verlust von Kompetenz***: Ein wesentlicher Aspekt, der auch hinter den vorangegangenen Risiken steht, ist der kurzfristige Verlust von Kompetenz, der sich darin manifestieren kann, dass Menschen immer stärker in die Abhängigkeit von Ai gelangen. Viel problematischer ist dabei auch die Abhängigkeit von denjenigen, die die Macht über die AI besitzen und diese kontrollieren (siehe unten). Kurzfristige Verluste von Kompetenz können jedoch, im Gegensatz zu langfristigen Kompetenzverlusten, recht schnell erkannt und mitigiert werden, daher wollen wir uns an dieser Stelle nicht zu stark mit ihnen beschäftigen.

- ***Langfristiger Verlust von Kompetenzlevels***: Viel diffiziler als der kurzfristige Verlust an Kompetenz ist die langfristige Reduktion des Talente-Pools und der Verlust fortgeschrittener Kompetenz-Leves, am besten beschrieben anhand der Aufgabe der Programmierung: Um einen Junior Developer zu ersetzen mag AI ausreichen, allerdings nach derzeitigem Stand wahrscheinlich nicht für einen Senior Developer, der auf Basis seiner Kenntnisse weitreichende Designentscheidungen treffen muss. Woher kommen jedoch die Senior Developer der Zukunft, wenn diese nicht als Junior Developer anfangen, ineffizient arbeiten und durch Fehler lernen können um dann mit diese Erfahrungen z u wachsen? Sollte es daher nicht möglich sein, auch die Senior Developer durch AI zu ersetzen und somit die gesamte Ausbildungskette ersetzbar zu machen, entsteht das Problem, das durch den Mangel an Juniors nach einigen Jahren keine Seniors mehr existieren. Dieses Problem ist umso diffiziler, da es erst zeitverzögert eintritt und am Anfang die kurzfristig performantere Strategie darstellt: Anstatt ineffizient mit menschlichen Juniors und deren Fehlern arbeiten zu müssen, arbeitet man effizient mit einem Gespann an AI und Seniors. Das Problem des „Wegsterbens“ der Seniors wird erst viel später wichtig, gerade in modernen Firmen mit ihren kurzen Zyklen der Führungsdauer auf dem C-Level erst in einigen Iterationen an CEOs, d.h. derjenige, der das Problem eigentlich verursacht wird mit perfekten Zahlen belohnt, während erst viel Jahre später die durch ihn gezeitigten Probleme schlagend werden. Dies ist für LEAs ein besonderes Problem, buhlen diese Institutionen doch üblicherweise um die Talente mit der Wirtschaft, ohne die dort möglichen Gehälter zahlen zu können. Sie sind daher von einer Knappheit an Expertise am meisten betroffen. Dieses Problem der Reduktion des Talente-Pools, bzw. der Senioren Level wurde von sehr vielen Interviewten erkannt und als eines der größten Probleme der modernen Ideen zum Einsatz von AI beschrieben.

- ***Haftungen***: Ein wesentlicher Aspekt, gerade vor der Herausforderung des Einsatzes von AI durch LEAs, ist die Frage der Haftung, speziell wenn die gegenständliche AI-Einfluss auf Ermittlungsverfahren hat. Für viele Aufgaben im Bereich des Law Enforcements gibt es strenge Regeln, die beachtet werden müssen, bspw. um Vorurteilsfreiheit zu erlangen oder um Befangenheiten zu vermeiden. Wie mit Bias in AI umzugehen ist, der direkten Einfluss bspw. auf Ermittlungen zeitigt, ist derzeit noch nicht klar und muss entsprechend geregelt werden. Allerdings ist anzunehmen, dass durch die Herstellerfirmen versucht wird, die Haftungsfragen auf den Einsetzend überzuwälzen. Viel wesentlicher ist allerdings die Frage, wie sichergestellt werden kann, dass diese Biases in den Trainingsdaten möglichst geringen Einfluss auf die polizeiliche Arbeit besitzen. Das Thema der Haftungen

ist auch in anderen Anwendungsfällen hochrelevant, speziell im Autonomen Fahren bei der Übergabe der Kontrolle von der Maschine an den Menschen wird die Frage aufgeworfen, ob damit auch die Haftung von der autonomen Steuerungssoftware auf den menschlichen Fahrer übergeht, dieser quasi in eine für ihn nicht handhabbare Situation geworfen wird, für die er dann auch noch haftet.

- ***Übermäßiges Vertrauen in AI***: Bereits angesprochen, aber wert noch einmal detaillierter beschrieben zu werden, ist die Gefahr des Verlusts des Misstrauens in maschinelle Entscheidungen bis hin zu einem Punkt, an dem keinerlei menschliche Kontrolle mehr stattfindet, sondern die Vorschläge der AI rein abgenickt werden. Dies kann so weit gehen, dass Menschen der AI grundsätzlich mehr vertrauen als ihren eigenen Sinnen und Erfahrungen und die AI damit quasi sogar die Vergangenheit „ändern“ kann, d.h. Menschen glauben eher der maschinellen Entscheidung und menschliche Kontrolle wird quasi nichtmehr durchgeführt.

- ***Verstärkung von Biases und Fehlern***: Durch die mangelhafte Kontrolle ist es auch vorstellbar, dass Biases (Vorurteile) in Daten und in den darauf aufbauenden Entscheidungen nicht mehr wahrgenommen und damit auch nicht mehr kontrolliert werden können, da den menschlichen Akteuren die Kompetenz zur Kontrolle fehlt. Dies würde u.U. nicht nur dazu führen, dass die Daten und darauf fußenden Entscheidungssysteme nicht besser im Sinne der Vorurteilsfreiheit werden würden, sie würden sogar Biases, die mit positivem Feedback bedacht werden, verstärken, d.h. es besteht die Gefahr, dass diese automatisierten Entscheidungssysteme immer mehr Vorurteile entwickeln. Diese neuen schlechten Daten würden im Gegenzug wieder zum Training anderer Algorithmen herangezogen, womit sich quasi ein richtiggehendes „Bias“-Ökosystem entwickeln könnte, in dem gewisse Vorurteile immer weiter verstärkt werden.

- ***Mangelhaftes Hinterfragen der Entscheidungs- und Optimierungsziele***: AI kann sehr schnell Entscheidungen treffen und im Fall guter Trainingsdaten kann die Qualität dieser Entscheidungen sehr gut sein. Dabei wird oft vergessen, was die Qualität einer AI-Entscheidung ausmacht, nämlich grundsätzlich die gute Lösung eines statistischen Optimierungsproblems, sei es das richtige Kategorisieren von Bildern in Kategorien, oder aber die Erstellung von Texten, die möglichst menschlich klingen. Dabei liegt die Problematik darin begründet, dass diese, im Folgenden vereinfacht als „Zielfunktion“ bezeichneten, Kriterien wesentlichen Einfluss auf die AI besitzen und überhaupt dazu geeignet sein müssen, ein gegebenes Problem zu lösen, bzw. die richtige AI mit der richtigen Zielfunktion für die Lösung einer Aufgabe genutzt werden muss. Trivial gesprochen sieht man dieses Problem schon heute an der Nutzung von ChatGPT: Nutzer*Innen beschweren sich, dass die Informationen in Texten von ChatGPT nicht korrekt sind, dabei war das gar nicht das ursprüngliche Design-Kriterium von ChatGPT, das darin lag, eine möglichst menschliche Unterhaltung / Informationsgenerierung zu imitieren. Auch Menschen sprechen nicht immer die Wahrheit, daher war das eigentlich kein primäres Ziel von ChatGPT. Auch im Bereich des Law-Enforcements kann es durch zu einfache oder sogar falsche Zielfunktionen zu Problemen kommen: Als Beispiel sie ein Algorithmus für das Predictive Policing gegeben, der Einsatzziele auf Basis bisheriger Verhaftungsdaten vorgibt und dessen Qualität in der Zahl der Verhaftungen definiert wurde, d.h. die Zielfunktion ist die Maximierung der Verhaftungsquote. Ein solcher Algorithmus würde bspw. Biases in den Daten 1:1 verstärken, da sie sein Ziel der maximalen Verhaftungsquote (zumindestens lokal) verbessern. Besitzt der Algorithmus umfangreichere Entscheidungsbefugnisse, würde er u.U. Präventionsprogramme reduzieren, da dieser Akt die Zahl der Kriminellen und damit auch die Zahl der

Verhaftungen erhöhen würde. Ein Problem an diesem System liegt direkt in der Zielfunktion begründet, das Ziel der polizeilichen Arbeit liegt zwar zu einem gewissen Teil darin, möglichst alle Kriminellen zu schnappen, aber das eigentliche Ziel (das jede/r Polizist*in implizit kennt) ist eigentlich die Reduktion von Verbrechen an sich, die Verhaftung ist lediglich ein Mittel zu diesem Zweck. Dies wurde aber in unserer Zielfunktion nicht vorgegeben. Diese Problematik, dass es de facto unmöglich ist Zielfunktionen seiteneffektfrei zu definieren ist derzeit auch ein wesentliches Thema einiger Forschungsprojekte, speziell aus dem Reinforcement Learning, bei dem sich das Problem noch potenziert.

- ***Machtkonzentration***: Wie bereits angedeutet, würde es der großflächige Verlust von Kompetenzen und die damit einhergehende Entmündigung vieler Menschen auch nach sich ziehen können, dass einige wenige, speziell denen, die die AIs und die zugrundeliegenden Datensätze kontrollieren, wesentliche Macht auf sich konzentrieren. Wir sehen die ersten Probleme der Machtkonzentration durch Kontrolle über Informationen und Informationstechnologien derzeit schon in der Politischen Landschaft, mit zunehmendem Kompetenzverlust und damit einhergehender Abhängigkeit von AI-Entscheidungen, aber auch durch die damit einhergehende erhöhte Schlagzahl in der Entscheidungsfindung, kann dieses Problem auch noch deutlich weiter eskalieren. Kontrollmechanismen müssten wiederum AI einsetzen, die mit entsprechenden Daten trainiert wurden. Eine Möglichkeit ist hierbei die Nutzung von sog. Controllable AI, die nicht versucht, Entscheidungen transparent nachrechnen zu können, sondern generelle Absicherungsmaßnahmen basierend auf Ergebnisanalyse und Analyse der Rohdaten vorzunehmen.

Einfluss auf menschliches Beziehungsverhalten: Ein weiterer wesentlicher Aspekt, der durch die Expert*Innen genannt wurde, war das Thema des Einflusses von AI-Nutzung auf zwischenmenschliches Verhalten. Dabei wurden zwei Kernaspekte prominent vertreten, die wir hier kurz wiedergeben wollen, auch, weil sie nach Meinung der Expert*Innen direkten Einfluss auf polizeiliche Tätigkeiten zeitigen können:

1. Die ***Unfähigkeit, zwischen menschlicher und KI-Kommunikation zu unterscheiden***: Mit zunehmender Qualität von AI-Systemen können diese auch aufgaben aus dem zwischenmenschlichen beriech übernehmen, bis hin zum Ersatz von Freund- und Partnerschaften. Speziell in der zwischenmenschlichen digitalen Kommunikation kann dies zu Problemen wie Vertrauensverlust führen, da man nichtmehr in der Lage sein wird zu unterscheiden, ob am anderen Ende gerade ein Mensch oder eine Maschine mit einem kommuniziert. Dieser Bruch auf einer persönlichen Ebene kann zu einem gesellschaftlichen Problem werden, bei dem auch andere Personen in der digitalen Kommunikation leichter herabgewürdigt und/oder entmenschlicht werden, einfach, weil man es gewohnt ist (oder es sich zumindestens einreden kann), dass am anderen Ende eine Maschine sitzt. Dies könnte zu einer weiteren Verrohung der Sprache und letztendlich des Denkens führen, was wiederum ein Problem des Law Enforcements darstellt, nicht nur in rein digitalen Bereichen zu Problematiken wie Hate Speech und/oder Mordaufrufen, sondern letztendlich auch im täglichen Leben zwischen Menschen. Ein weiterer Aspekt ist das Problem, dass umgekehrt die Beziehung zu AI eine attraktive Option für viele Menschen darstellen könnte, da diese KI sie bestätigt, ihre Interessen reflektiert und letztendlich einfacher und angenehmer im Umgang sein könnte als ein menschlicher Beziehungspartner. Dabei kann der Mensch u.U. rein kognitiv zwischen AI und anderen Menschen unterscheiden, will dies jedoch einfach nicht mehr. Dies könnte gravierende Auswirkungen auf die Gesellschaft

haben, die weit über Effekte auf die reine polizeiliche Tätigkeit hinausgehen, bis hin zu Problemen der Gesellschaftsspaltung, der Vereinsamung und letztendlich des Verlustes des sozialen Kitts im gesellschaftlichen Gefüge. Staatliche Akteure könnten diese Probleme gezielt verstärken, um Filterblasen zu erzeugen, in denen Hass auf gewiss Gruppen en vouge und akzeptiert ist, diese Problematiken würden sich auch in die Realität der polizeilichen Tätigkeit ziehen.

2. ***Effekte reiner M2M-Kommunikation auf die Gesellschaft***: Eng damit verwandt ist eine andre Beobachtung der interviewten Expert*Innen: AI-gestützte Kommunikation ermöglicht auch die Automatisierung gewisser Teile der Nachrichtenerstellung. Dies würde zu interessanten Effekten führen, wie bspw., dass der Sender seine AI bittet, aus fünf Stichwörtern ein nettes Mail zu entwerfen und dieses an einen Geschäftspartner zu schicken. Dieser wiederum bittet seine AI, das Mail (das er nicht selbst lesen will) in Stichwörtern zusammenzufassen – de facto spricht AI mit AI, wenn als Service ausgelagert u.U. sogar die gleiche AI, die auch direkt auf die Quellnachricht zugreifen kann. Dies könnte man weiterspinnen, dass dadurch (laut Expert*Innenmeinung) speziell Freundlichkeiten und höfliche Umgangsformen stark entwertet werden, zum einem, da sie durch die Zusammenfassung gefiltert werden, zum anderen, weil sie üblicherweise sowieso durch eine AI erstellt wurde – u.U. in Verschneidung mit Social Media könnten sogar sehr persönliche Aspekte dieser neuen, automatisierten Pseudohöflichkeit, wie bspw. Fragen zum Wohlergehen der Kinder, automatisiert erstellt werden. Dadurch ist aber der Wert solcher Fragen gleich Null, da sie nicht mehr echtes Interesse des Partners reflektieren, der noch dazu Zeit aufwendete, um sich zu erkundigen. Dies könnte laut unseren Expert*Innen zu einer Verrohung der Gesellschaft führen, da Höflichkeit grundsätzlich viel eher als Fake eingestuft wird. Auch dies wiederum könnte nicht nur die Umgangsformen mit der Polizei stören, sondern auch Hate Crime und zunehmender Entmenschlichung Vorschub leisten.

Grundsätzlich ist die Frage der Effekte von AI auf menschliches Beziehungsverhalten ein wichtiges offenes Forschungsgebiet der Sozialwissenschaften, dem speziell auch durch polizeiliche Akteure Aufmerksamkeit gewidmet werden sollte, da Delikte wie Hate Crimes und Hate Speech direkt dadurch betroffen sein und auch die Art der Kommunikation zwischen Polizisten und Bürger beeinflusst werden könnte.

Generierte Szenarien

Wir haben basierend auf den Scores, speziell aber auch in Hinblick auf häufig getroffene Aussagen aus den Interviews und konsistente zusätzliche Annahmen fünf Szenarien entwickelt, die mögliche Zukünfte beschreiben, Wir sind dabei nicht sklavisch dem Scoring gefolgt, da dieses grundsätzlich sehr stark gleichartige Szenarien bevorzugt, die sich lediglich in sehr geringem Ausmaß unterscheiden. Uns war es wichtig, zum einen Möglichkeiten, die in den Interviews avisiert wurden, auch konkret darzustellen, auf der anderen Seite auch eine gewisse Bandbreite an möglichen Szenarien zu skizzieren.

Um einen weiteren Mehrwert zu generieren, haben wir nicht nur die ausgewählten Faktorausprägungen dargelegt, sondern auch Gaps/Herausforderungen und Effekte auf LEAs, sowie neue möglich Anwendungen im Umfeld der LEAs beschrieben. Dabei kann es sich natürlich niemals um vollständige Aufzählungen handeln, zum einen, da selbst kleine Abweichungen die Möglichkeiten neuer Anwendungen stark beeinflussen, zum anderen, da eine Vollständigkeit implizieren würde,

dass wir in der Lage wären, alle guten Geschäftsideen in diesem Umfeld zu extrahieren, eine Aufgabe, die trivialerweise nicht erwartet werden kann.

Die Szenarienbeschreibungen selbst wurden eher kurz gehalten und entsprechend den Regeln des Storytellings aus dem Originalansatz von Geschka umschrieben ausgeführt. Durch die Nutzung eines LLMs kann dies natürlich noch erweitert und weiter ausgeschmückt werden, dies haben wir hierbei allerdings dem/r Leser*In überlassen, nicht zuletzt aufgrund möglicher Copyright-Problematiken bei der Texterstellung durch einen GPT.

Disclaimer: Es ist wesentlich zu verstehen, dass, obwohl die Szenariengenerierung auf den Aussagen von Expertinnen in ihren jeweiligen Fachgebieten basieren, dies keine exakte Vorhersage sein kann, d.h. (i) hängt die Szenariengenerierung sehr wesentlich davon ab, was Expert*innenzu einem gewissen Zeitpunkt zu einem Thema eingefallen ist und (ii) kann keine expert*in die Zukunft vorhersagen. (iii) ist die Methode der ESA selbst nicht ausreichend exakt, um eindeutige Szenarien aus den (unscharfen) Aussagen aus Interviews zu erzeugen, zusätzlich (iv) wird bei der Beschreibung der Szenarien basierend auf den Extrapolationen immer eine willkürliche Ausschmückung, aber auch Weglassung betrieben. Es ist daher wesentlich, dass die Szenarien nur Beispiele für eine sinnvolle und (hoffentlich) konsistente Zukunftsentwicklung darstellen, die Zukunft sich aber natürlich auch anders entwickeln kann.

Szenario 1 – „Mittelweg“

Ausgewählte Faktorausprägungen:

Für dieses Szenario wurde eine Kombination der folgenden Faktorausprägungen ausgewählt. Dabei handelte es sich um ein Szenario mit einem sehr hohen Scoring, d.h. viele der Faktorausprägungen wurden von mehreren Expert*Innen genannt:

- Regulation von AI findet einen Mittelweg zwischen Innovationskraft und Schutz
- EU schafft eine eigene AI-Industrie, wesentliche Tools können kontrolliert werden
- Kosten von AI durch Algorithmen und Hardware für sehr viele Anwendungen ausreichend gering
- Qualität ist für die meisten Branchen ausreichend erzielbar
- Auswirkungen auf den Arbeitsmarkt Nullsummenspiel, aber einige Abgehängte
- Jobs in Verwaltung und Low-level-Coding verschwinden
- AI ist sehr populär und allgegenwärtig, Einsatz und Kontrolle meist transparent
- Viele Personen verstehen AI-Grundlagen
- Persönliche AI-Systeme sind auf Personen zugeschnitten und respektieren die Privatsphäre
- Private sharen ihre Daten sehr freizügig auf Social Media und mit AI-Services im Internet
- Sharing Economy boomt, daraus folgt die gemeinsame Nutzung von Devices in verschiedenen kommerziellen und nichtkommerziellen Modellen
- LEAs bekommen ausreichend und schnell AI-Werkzeuge
- Einsatz von AI bei LEAs wird stark ge-monitored
- Geopolitische Stabilität ist gegeben

Szenariobeschreibung

Dieses Szenario bildet einen Mittelweg zwischen Szenario 2 „Abstieg“ und Szenario 4 „Wild AI West“. In vielen Aspekten bildet dieses Szenario tatsächlich einen Mittelweg. Regularien zum Schutz von Personen sind vorhanden, aber behindern die Entwicklung von AI und AI-Anwendungen nicht zu sehr, wodurch auch in Europa eine gute und schlagkräftige AI-Industrie entsteht. Dazu trägt bei, dass die Kosten für AI weiter in einem guten Ausmaß sinken und für die meisten Anwendungen eine ausreichende Qualität günstig erzielbar ist. Natürlich verschwinden Arbeitsplätze durch den weitereichenden Einsatz von AI, was zu einer gewissen Schicht an Unzufriedenen und Abgehängten führt, aber es entstehen gleichzeitig genügend gute Jobs, dass dies durch das Sozialsystem und Umschulungsmaßnahmen abgefangen werden kann. Diese verschwindenden Berufe betreffen hauptsächlich rein organisatorische Arbeiten, sowie Low-Level-Coding.

Durch die Popularität von AI beding verstehen auch viele Personen die Basisgrundlagen, speziell, was sie von AI erwarten könne und wo die Gefahren liegen. AI wird stark personalisiert, Regularien sorgen dafür, dass die Privatsphäre durch die Tools gut geschützt und respektiert wird, allerdings befinden wir uns in einer starken Sharing Economy, in der viele Leute freizügig mit Devices, aber auch ihren privaten Daten umgehen. Auf Social Media aktiv zu sein, wird sehr beliebt und auch AI-Services im Internet werden mit privaten Daten gefüttert.

LEAs bekommen grundsätzlich sehr rasch gute und neue Tools, die sie bei ihren Aufgaben unterstützen, allerdings findet ein sehr starkes Monitoring bei der Nutzung statt, speziell um mögliche missbräuchliche Verwendung rasch aufdecken und sanktionieren zu könne, was auch im Regelfall geschieht. Wir befinden uns in einer Phase wiedergewonnener geopolitischer Stabilität mit einem ausreichend starken Europa.

Generelle Effekte auf LEAs

Folgen eine Skizzierung der möglichen extrapolierten Effekte auf Law Enforcement Agencies, basierend auf der Gestalt des vorliegenden Szenarios, zusammen mit kurzen Anmerkungen und Beschreibungen, wo dies für notwendig erachtet wurde.

- Guter Talente-Pool aus dem die LEAs schöpfen können, ermöglicht es auch kompetitiv am Markt Expert*Innen zu interessieren und entsprechendes Personal aufzubauen.
- Vertrauenswürdige und günstige Werkzeuge ermöglichen den Einsatz von AI für viele Anwendungen und viele Nutzungspotenziale.
- AI-Einsatz wird alltäglich, daher ist auch die Übung, die selbst Nichtspezialist*Innen in der Polizei besitzen, hoch
- Viele Low-Level-Angriffe verfangen nicht, da auch normale Bürger typischerweise in der Lage sind, diese zu erkenne, speziell im Bereich der Fake News und Grassrooting, sowie spezifisch im Bereich des Phishings mit AI.
- Verbesserte Aufklärung (spez. OSINT) durch LEAs, zum einen durch das Vorhandensein sehr guter Werkzeuge, zum anderen auch durch die Freizügigkeit der Daten und Informationsweitergabe von Privaten in allen möglichen Kanälen, speziell auch Social Media.
- Schlechte Attribution von Angriffen, speziell durch die starke Nutzung gemeinsamer Ressourcen im Rahmen der Sharing Economy nutzen Kriminelle präparierte gesharte Geräte und Devices zur Verdeckung des tatsächlichen Ausgangspunktes von Angriffen.

Kriminelle können auf eine Vielzahl an gesharten Geräten zugreifen und sich diese nutzbar machen. Attack Attribution ist damit extrem schwierig.

- Datenquellen sind zuverlässig, da sie durch den Aufbau einer eigenen Industrie bedingt viel Aufmerksamkeit erhielten, die Problematik von Fremdkontrolle über Datenquellen daher sehr früh erkannt und mitigiert wurde. Wo fremdbestimmte Datenquellen notwendig sind, sind diese mit Hilfe von Techniken wie Controlable AI und dergleichen speziell abgesichert und werden regelmäßig hinterfragt.
- Nutzung von AI in der Bevölkerung akzeptiert, speziell auch durch LEAs, da es der/dem normalen Bürger*In klar ist, dass dieses für sie selbst selbstverständliche Werkzeug auch in diesen Bereichen des Lebens genutzt werden. Die starke Kontrolle in Bezug auf Missbrauch verstärkt diese Akzeptanz.
- Kriminelle analysieren LEA-Verhalten: Durch die weite Verbreitung von AI-Wissen stehen auch Kriminellen viele Werkzeuge und viel Wissen im Bereich der Nutzung von AI in Analysen zur Verfügung. Dieses nutzen sie, um Muster in der Polizeiarbeit zu erkennen, bspw. Rückschlüsse auf OSINT-Strategien oder Ressourcen-Allokation im polizeilichen Dienst zu ziehen. Dabei greifen sie nicht direkt die AI-Werkzeuge der Polizei an, sondern sie nutzen lediglich eigene AI um dieses Wissen über Muster und Strategien zu erkennen.
- Attack-Surface AI in LEAs: Die Integration neuer Werkzeuge birgt auch immer wieder sicherheitstechnische Herausforderungen, speziell wenn Non-Explainable AI zu Back Boxes führt, die nur unzureichend durchschaut und durchdrungen werden können. Angreifer werden versuchen, die Werkzeuge der Polizei anzugreifen, allerdings gehen wir in diesem Szenario davon aus, dass die gute Kontrolle gegen missbräuchliche Verwendung als Nebeneffekt quasi auch sehr gute Absicherung gegen Angreifer von Innen und Außen impliziert.

Neue Anwendungen für LEAs

Folgend geben wir einen kurzen Überblick darüber, welche neuen Anwendungen wir in dem gegenständlichen Szenario als sehr wahrscheinlich erachten. Eine Beschreibung findet sich im Anschluss an die Szenarienbeschreibung.

- Information-/Model-Sharing mit anderen (internationalen) Entitäten
- Daten-Integration mit anderen staatlichen Stellen
- Gemeinsames Trainings-Center
- Einbindung privater Sensorik/Geräte als Informationsquelle für hohe Informationsdichte
- Scenario Forecasting
- Persönliche KI-Assistenten
- Analysewerkzeuge für den polizeilichen Dienst
- AI für die Ressourcenallokation
- Automatisierung von Basisarbeiten und Managementtätigkeiten

Lücken und resultierende Herausforderungen

Folgend geben wir einen kurzen Überblick darüber, welche neuen Lücken, Gefahren und Herausforderungen wir in dem gegenständlichen Szenario als sehr wahrscheinlich erachten. Eine Beschreibung findet sich im Anschluss an die Szenarienbeschreibung.

- Austrocknung des Talente-Pools
- Extreme Low-Tech-Crimes erleben Renaissance
- Sehr große Datenmengen mit viel Noise
- Viele aktive Cyber-Kriminelle,
- Problematische neue Angriffe via shared Devices
- Mangelhafte Kontrolle von Einzelnen in LEAs
- Unbeabsichtigte Informationsweitergabe
- Absicherung gegen Re-Extraktion von Trainingsdaten
- Weitergabe strategischer Meta-Informationen durch Szenariengestaltung
- Proaktive Erkennung privater Informationen schwierig
- Absicherung gegen Weitergabe der gesharten Daten
- Absicherung gegen Backdoors in pre-trained models
- Private Systeme werden u.U. Angriffsziel
- Angriffe durch präparierte gesharte Geräte
- Attack- und Backdoor-Attribution schwierig
- AI-generiertes dezentrales data flooding erkennen
- Fake-News und Information Warfare zielt auf Abgehängte ab
- Angriffsoberfläche im Bereich der LEAs wächst
- Blind Spots durch Analyse der LEA-Werkzeuge

Szenario 2 – „Abstieg“

Ausgewählte Faktorausprägungen:

Für dieses Szenario wurde eine Kombination der folgenden Faktorausprägungen ausgewählt. Dabei handelte es sich um ein Szenario, das in gewisser Weise einen dualen Charakter zum ersten Szenario aufweist, indem einige dort sehr positive Aspekte nicht eintreffen. Dieses Szenario fußt auch auf den Befürchtungen einiger Interviewpartner*Innen für die zukünftige Entwicklung.

- Regulation von AI ist zu stark
- Keine relevante AI-Industrie in der EU
- Kosten von eigener AI sehr hoch, totale Abhängigkeit von anderen Staaten
- Qualität ist für die meisten Branchen als „AI as a Service“ (AIaaS) ausreichend erzielbar
- High-Level-Jobs in IT verschwinden
- AI wird im Rahmen von Internet-Services permanent genutzt, Einsatz und Kontrolle intransparent
- Kaum jemand außerhalb einer gewissen Tech-Clique versteht AI-Grundlagen, rein passive Nutzung dominiert
- Persönliche AI-Systeme existieren kaum, Bezahlung von AI-Services (online) mit persönlichen Daten ist üblich
- Private teilen ihre Daten sehr freizügig auf Social Media und mit AI-Services
- Kaum ausgebildete Sharing-Economy im digitalen Bereich
- LEAs bekommen kaum AI-Werkzeuge
- Geopolitische Stabilität ist gegeben

Szenariobeschreibung

Dieses Szenario ist in gewisser Weise in vielen Aspekten ein duales Szenario zum ersten Szenario, mit der Einschränkung, dass in beiden Fällen eine grundsätzliche geopolitische Stabilität gegeben ist.

Ai ist in der Europäischen Union sehr stark reguliert, es konnte sich keine eigene Ai-Industrie herausbilden, weder auf Seite der Grundlagen und grundlegenden Techniken noch auf Seite der Anwendungen. AI-Services werden von Bürger*Innen rege genutzt, diese sind jedoch ausschließlich n den Händen der großen AI-Player, vornehmlich aus den USA und China. Die AI-Grundlagen und Gefahrenpotenziale werden von den meisten Anwender*Innen nicht verstanden und auch nicht bedacht. AI ist für viele Anwender*Innen ein Online-Tool wie vieles andere, über das kaum nachgedacht wird. Bezahlung mit privaten Daten ist üblich und wird kaum hinterfragt auch nicht in Bezug auf Gefahren. Dies ist sehr konsistent mit dem Trend, grundsätzlich Privates auch auf vielen anderen Plattformen zu teilen. Im Gegensatz dazu nehmen wir an, dass es zu keiner Herausbildung einer echten Sharing-Economy kommt, d.h. abgesehen davon, dass fremdkontrollierte Dienste mit privaten Daten quasi bezahlt werden, ist kein echtes Sharing üblich.

Eigene AI-Komponenten und Systeme sind in der EU sehr teuer und werden daher kaum genutzt, daher werden auch im Umfeld der Industrie Services aus USA/China genutzt, da der Aufbau eigener Services in den meisten Fällen prohibitiv teuer ist. Dadurch verschwinden auch stetig die wirklich innovativen Jobs und Personen aus der IT-Branche, aber auch aus vielen anderen Branchen.

Dies führt auch dazu, dass LEAs kaum AI-Werkzeuge bekommen, da diese, wenn nicht als Service eingekauft, zu teuer sind. Auch der Talente-Pool trocknet aus, die guten Leute gehen ins Ausland, es bleiben kaum sinnvoll in AI ausgebildete Personen für den polizeilichen Dienst übrig. Dem gegenüber stehen Kriminelle, die durchaus entsprechende Kapazitäten aufbauen und daher entsprechend agieren können, bspw. um Aktionen der Polizei vorhersagen und generell ihre technologische Überlegenheit auszuspielen.

Generelle Effekte auf LEAs

Folgen eine Skizzierung der möglichen extrapolierten Effekte auf Law Enforcement Agencies, basierend auf der Gestalt des vorliegenden Szenarios, zusammen mit kurzen Anmerkungen und Beschreibungen, wo dies für notwendig erachtet wurde.

- Schlechter Talente-Pool, da technische gute Leute ins Ausland, bevorzugt Länder mit viel Ai.-Kompetenz, abwandern. Die, die bleiben gehören einer kleinen technischen Elite an, für die die Arbeit in LEAs typischerweise nicht interessant ist. Einige technisch versierte Leute werden auch vom organisierten Verbrechen bezahlt, da diese vermehrt auf die Nutzung von Ai gegen die Polizei setzen.

- Werkzeuge und Modelle sind veraltet und unbrauchbar, da eine Eigenentwicklung für LEA viel zu teuer und aufwändig ist, speziell kaum Wissen innerhalb der LEAs existiert, Dies gilt auch für Tätigkeiten wie das Training der Modelle und dergleichen. Services werden in diesem Szenario von LEAs nicht, oder nur sehr reduziert, in Anspruch genommen, da diese rein fremdbestimmt durch andere Staatensind und sich daher nicht für die polizeiliche Arbeit eignen. Dies kann natürlich, je nach wechselnder politscher Lage, diskutiert werden.

- AI-Einsatz wird Feigenblatt: Sollte AI eingesetzt werden, dann vermehrt, um politischen Wilen zeigen zu können und den/die Bürg*Innen damit zu beeindrucken und zu beruhigen,

dass natürlicherweise moderne Technik in der Strafverfolgung eingesetzt wird. Allerdings zeitigt dieser Einsatz kaum echten Nutzen.

- Hoher krimineller Druck, da Cyber-Kriminelle aus dem In- und Ausland verstärkt auf AI setzen, auch auf Technologien, die in anderen Staaten bereits chancenlos veraltet sind. Damit nutzen sie auch das Problem der unwissenden Bevölkerung, sowie das Vorhandensein vieler privater Daten und Informationen, die für eine Vielzahl an Angriffen, beginnend bei etwas besserem Phishing, bis hin zu komplexen CEO-Frauds genutzt werden kann.
- Cyber-Kriminelle haben sehr geringen Einstiegslevel, da bereits die Nutzung sehr einfacher, vorgefertigter Tools ausreichend ist, um einen gewissen Grad an Erfolg zu erreichen. Dies führt dazu, dass sich auch die Zahl der Cyber-Kriminellen drastisch erhöht, die Gefahr erwischt zu werden sinkt und das für erfolgreiche Angriffe notwenige Wissen kann sehr rasch angeeignet werden.
- Niedrige Aufklärungsquote, da die genutzten Tools stark veraltete sind, Kriminelle gleichzeitig der Polizei haushoch überlegen sind, was die Nutzung moderner Technologie angeht.
- Nutzung von AI in der Bevölkerung kaum akzeptiert: Da die Bevölkerung AI nur als Online-Service nutzt und die Entwicklung eigener Software nicht üblich und sehr teuer ist, ist die Nutzung durch LEAs auch in der Bevölkerung nicht sehr akzeptiert. Allerdings spielen Privacy-Gründe weniger eine Rolle bei dieser Abneigung.
- Kriminelle besitzen massiven Vorsprung in der Nutzung von AI, speziell wenn sie aus anderen Staaten aus agieren, oder sogar von diesen Unterstützung erhalten. Dies führt dazu, das Kriminelle AI auch gegen die Polizei einsetzen, bspw. Zur Prediction von Verhalten und Ressourcenallokationen, aber auch ganz banal zur Automatisierung von Angriffen.

Neue Anwendungen für LEAs

In diesem Szenario rechnen wir nicht mit nennenswerten neuen Anwendungen von AI für LEAs, speziell, da die Anforderung der Kontrolle über solche Werkzeuge nicht konsistent mit den prohibitiv hohen Kosten ist.

Lücken und resultierende Herausforderungen

Folgend geben wir einen kurzen Überblick darüber, welche neuen Lücken, Gefahren und Herausforderungen wir in dem gegenständlichen Szenario als sehr wahrscheinlich erachten. Eine Beschreibung findet sich im Anschluss an die Szenarienbeschreibung.

- Cyber-Crime floriert, auch Low-Level-Phishing mit AI hat viel Erfolg
- Starke Zunahme von Impersonation Attacks und Identity Theft
- High-Level Attacken und AI-basierte Botnetze florieren
- Polizei wird predictable, d.h. das organisierte Verbrechen nutzt AI aktiv gegen die Polizeiarbeit.
- Attack-Attribution wird unmöglich, speziell im Bereich der Impersonation, da auf der einen Seite sehr viele Daten frei verfügbar sind, auf der anderen Seite die Möglichkeiten der

Polizei zu stark eingeschränkt wurden, auch wenn keine echte Sharing-Economy mit ihren negativen Effekten auf die Attribution vorhanden ist.
- Data-Flooding, d.h. es entstehen große Mengen an potenziell interessanten Daten durch die Tätigkeit der Bürger*Innen, diese können jedoch nicht ausgewertet werden.
- Viel Fake-News und Fremdbeeinflussung (Information Warfare)
- Keine Kontrolle über Informationskorrektheit

Szenario 3 – „Passiv Aktiv“

Ausgewählte Faktorausprägungen:

Für dieses Szenario wurde eine Kombination der folgenden Faktorausprägungen ausgewählt. Das Szenario ist in vielen Aspekten eine Abschwächung des zweiten Szenarios und unterscheidet sich hauptsächlich in der Stellung von AI für LEAs.

- Regulation von AI in der EU sehr stark
- Kein Aufbau einer eigenen AI-Industrie
- High-Level-Jobs in IT verschwinden
- Gute Werkzeuge als AIaaS zur Verfügung
- AI ist sehr populär und allgegenwärtig, Einsatz und Kontrolle extrem intransparent
- Viele Personen verstehen AI-Grundlagen
- Persönliche AI-Systeme sind auf Personen zugeschnitten
- Bezahlt für AI-Services wird oft mit privaten Daten
- Private sharen ihre Daten sehr freizügig auf SM und mit AI
- Jobs in Verwaltung und Low-level-Coding verschwinden
- Qualität ist für die meisten Branchen ausreichend erzielbar
- Kosten als AIaaS sehr gering
- LEAs haben Zugriff auf sehr neue Werkzeuge, komplett außerhalb ihrer Kontrolle
- Geopolitische Stabilität ist gegeben

Szenariobeschreibung

Dieses Szenario ist in vielen Aspekten sehr ähnlich zu dem zweiten Szenario, beschreibt allerdings einen weniger harten Abstieg, sondern eine sehr passive Verhaltensweise gegenüber Ai in der Bevölkerung, mit gleichzeitig sehr großer Nutzung durch LEAs.

Auch hierbei führt die starke Regulierung von AI und deren Nutzung in der Europäischen Union dazu, dass keine eigenen starke AI-Industrie aufgebaut wird, allerdings stehen gute Werkzeuge als fremdbestimmte Online-Services zur Verfügung. Ai ist in der Nutzung sehr populär und wird auch durch sehr viele Personen fast täglich genutzt, oft auch ohne dies konkret zu wissen, d.h. der Einsatz ist sehr intransparent, Kontrolle über die Nutzung der Services, und was sie tatsächlich tun gibt es kaum. Allerdings sind viele Bürger*Innen durchaus mit den Grundlagen von AI vertraut, was ihnen recht gute Chancen gegen Low-Level-Angriffe wie AI-basiertes Phishing und dergleichen gibt, die verbreitet, aber weniger ein gesellschaftliches Problem darstellen. Ansonsten ist der wesentliche Unterschied zum zweiten Szenario in der Nutzung von AI durch LEAs zu sehen, diese bekommen durch den Staat guten Zugriff auf sehr neue Werkzeuge, die allerdings komplett außerhalb ihrer Kontrolle liegen. Dennoch kann mit Hilfe dieser Werkzeuge effektiv gegen

Kriminelle vorgegangen werden, die regelmäßigen Updates reduzieren auch die Gefahr, dass diese Werkzeuge als Angriffsvektoren gegen die LEAs verwendet werden.

Generelle Effekte auf LEAs

Folgen eine Skizzierung der möglichen extrapolierten Effekte auf Law Enforcement Agencies, basierend auf der Gestalt des vorliegenden Szenarios, zusammen mit kurzen Anmerkungen und Beschreibungen, wo dies für notwendig erachtet wurde.

- Mittelmäßiger Talente-Pool: Auf der einen Seite beherrschen viele Bürger*Innen die Grundlagen der AI und verstehen auch mit AI umzugehen, was sie als Grundstock für die Ausbildung on diesem Bereich sehr begehrenswert macht und grundsätzlich auch den Talente-Pool bereichert. Auf der anderen Seite steht das Problem, dass echte Spezialist*Innen wenig Attraktivität im heimischen Arbeitsmarkt sehen und in das AI-affinere Ausland mit eigener AI-Industrie und entsprechenden Möglichkeiten abwandern. Es ist daher fraglich, ob man auf Basis des guten Grundstockes selbst geeignetes Personal ausbilden kann und wie lange dies gehalten werden kann.

- Günstige Werkzeuge: Günstige Werkzeuge ermöglichen den Einsatz von AI für viele Anwendungen und viele Nutzungspotenziale, allerdings ist die technische Kontrolle über diese Tools nicht immer gegeben, speziell da sie extern eingekauft werden müssen. Hierbei helfen u.U. die starken Regularien, die entsprechende Kontrollmechanismen vorschreiben könnten, die dadurch die Vertrauenswürdigkeit erhöhen, sollte eine entsprechende Einkaufspolitik verfolgt werden.

- AI-Einsatz wird alltäglich, allerdings ist er oft sehr passiv, da die Tools lediglich bedient werden und der Talente-Pool überschaubar ist.

- Viele Low-Level-Angriffe verfangen nicht, da auch normale Bürger typischerweise in der Lage sind, diese zu erkenne, speziell im Bereich der Fake News und Grassrooting, sowie spezifisch im Bereich des Phishings mit AI.

- Verbesserte Aufklärung (spez. OSINT) durch LEAs, zum einen durch das Vorhandensein sehr guter Werkzeuge, zum anderen auch durch die Freizügigkeit der Daten und Informationsweitergabe von Privaten in allen möglichen Kanälen, speziell auch Social Media.

- Zuverlässigkeit der Datenquellen: Grundsätzlich gehen wir davon aus, dass ein Grundstock an Daten aus eigenen Analysen auf Basis der erweiterten OSINT-Fähigkeiten stammen und diese auch durch die LEAs verstanden werden und entsprechend qualitätsgesichert werden können. Allerdings stellt sich die Frage der Modelle und ob diese in diesem Szenario wirklich von den LEAs selbst trainiert werden könne, hierzu müssten also entsprechende Kontrollmechanismen geschaffen werden, die bei Training durch Dritte schlagend werden. Durch das Fehlen der eigenen Ai-Industrie ist die technische Umsetzung dieser Kontrollen fraglich, wir sehen auch das Problem, dass eine große Abhängigkeit von externen Modellen, Trainings und letztendlich auch Daten in diesem Szenario mit großer Wahrscheinlichkeit schlagend werden wird. Grundsätzlich gehen wir allerdings im Szenario von einer Periode großer geopolitischer Stabilität aus, d.h. es wird davon ausgegangen, dass keine direkten Feindseligkeiten „befreundeter“ Regierungen stattfinden, die die Nutzung von AI als Services durch LEAs grundsätzlich verunmöglichen.

- Nutzung von AI in der Bevölkerung je nach Anwendung umstritten: Wir nehmen in diesem Szenario an, dass die Bevölkerung genügend über AI weiß, um sich der bisher genannten Problematiken des großen Kontrollverlusts der LEAs über ihre Systeme als rein passive Nutzer und den daraus resultierenden Risiken bewusst sind und daher dem Einsatz von AI, speziell in Hinblick auf Prozesse mit politischer Komponente, kritisch gegenüberstehen, d.h. obwohl die echte Expertise zum Aufbaueiner eigenen AI-Industrie fehlt, ist das Bewusstsein der Bevölkerung durch Basiswissen, aber auch durch Diskussionen um Regularien hinreichend geschärft. Wir sehen daher u.U. Probleme bei der Akzeptanz der Nutzung von AI in Prozessen und im Rahmen der Beweismittelfindung gegeben.

- Kriminelle analysieren LEA-Verhalten: Durch die weite Verbreitung von AI-Wissen stehen auch Kriminellen viele Werkzeuge und viel Wissen im Bereich der Nutzung von AI in Analysen zur Verfügung. Dieses nutzen sie, um Muster in der Polizeiarbeit zu erkennen, bspw. Rückschlüsse auf OSINT-Strategien oder Ressourcen-Allokation im polizeilichen Dienst zu ziehen. Dabei greifen sie nicht direkt die AI-Werkzeuge der Polizei an, sondern sie nutzen lediglich eigene AI um dieses Wissen über Muster und Strategien zu erkennen.

Neue Anwendungen für LEAs

Folgend geben wir einen kurzen Überblick darüber, welche neuen Anwendungen wir in dem gegenständlichen Szenario als sehr wahrscheinlich erachten. Eine Beschreibung findet sich im Anschluss an die Szenarienbeschreibung.

- Information-/Model-Sharing mit anderen (internationalen) Entitäten
- Bezahlung mit Modellen und Feedback
- Scenario Forecasting
- Analysewerkzeuge für den polizeilichen Dienst
- AI für die Ressourcenallokation
- Automatisierung von Basisarbeiten und Managementtätigkeiten

Lücken und resultierende Herausforderungen

Folgend geben wir einen kurzen Überblick darüber, welche neuen Lücken, Gefahren und Herausforderungen wir in dem gegenständlichen Szenario als sehr wahrscheinlich erachten. Eine Beschreibung findet sich im Anschluss an die Szenarienbeschreibung.

- Austrocknung des Talente-Pools
- Low-Tech-Crimes erleben Renaissance
- Fehlende Kontrolle über LEA-Daten
- Fehlende Kontrolle über LEA-Modelle
- Fehlende Kontrolle über Auswertungsergebnisse
- Privacy fast nicht mehr durchsetzbar
- Sehr große Datenmengen mit viel Noise
- Kontrollmöglichkeiten, Nachvollziehbarkeit und Transparenz hängen vom Tool-Anbieter ab
- Unbeabsichtigte Informationsweitergabe
- Absicherung gegen Re-Extraktion von Trainingsdaten
- Mangelhafte Qualität der Trainingsdaten
- Schaffung von Nachvollziehbarkeit und Reproduzierbarkeit

- Marketing für AI für LEAs
- Rating und Monitoring der AI-Tools für LEAs
- Absicherung gegen Backdoors in pre-trained models
- Angriffsoberfläche im Bereich der LEAs wächst
- Blind Spots durch Analyse der LEA-Werkzeuge

Szenario 4 – „Wild AI West“

Ausgewählte Faktorausprägungen:

Für dieses Szenario wurde eine Kombination der folgenden Faktorausprägungen ausgewählt. Das Szenario ist in vielen Aspekten eine Abschwächung des zweiten Szenarios und unterscheidet sich hauptsächlich in der Stellung von AI für LEAs.

- Regularien werden deutlich abgeschwächt bis hin zur fast vollständigen Aufhebung
- EU schafft eine kompetitive AI-Industrie, wesentliche Tools können kontrolliert werden
- Sehr großer Talente-Pool
- Kosten von AI durch Algorithmen und Hardware für sehr viele Anwendungen ausreichend gering
- Qualität ist für die meisten Branchen ausreichend erzielbar
- „Es wird gemacht, was geht“ – Kaum Monitoring, keine Kontrolle, kaum echte Nachvollziehbarkeit
- Positive Auswirkungen auf den Arbeitsmarkt, aber einige Abgehängte
- Jobs in Verwaltung und Low-level-Coding verschwinden rasch
- AI ist sehr populär und allgegenwärtig, Einsatz und Kontrolle komplett intransparent
- Viele Personen verstehen AI-Grundlagen
- Persönliche AI-Systeme sind auf Personen zugeschnitten, Bezahlen mit Daten normal
- Private sharen ihre Daten sehr freizügig auf SM und mit AI
- LEAs bekommen sehr schnell neue AI-Werkzeuge
- Einsatz von AI bei LEAs wird nicht kontrolliert, qualitätsgesichert oder nachvollziehbar gehalten
- Geopolitische Stabilität ist gegeben

Szenariobeschreibung

In diesem Szenario werden in Bezug auf AI und deren Anwendung kaum mehr Barrieren aufgestellt, speziell im Bereich der Regularien wird extrem aufgeweicht, nicht nur direkt AI betreffende Regelungen wie den AI-Act betrifft, sondern auch indirekt in Bezug auf Datenschutz und Datensicherheit. Es zählt die Innovation und die Geschwindigkeit dieser, Schäden werden erst nach Implementierung und oftmals Ausrollen von Technologie analysiert und ggf. wird nachträglich eine gewisse Regelung konkreter auftretender Probleme in Betracht gezogen. Dies ermöglicht den Aufbau einer immens schnellen und unregulierten AI-Industrie mit extrem vielen Anwendungsgebieten. Aber auch einer gewissen Unsicherheit, weil grundsätzliche Guidelines fehlen und Anwendungen erst nachträglich im Schadensfall reguliert werden, was dazu führt, dass sich auch Firmen etablieren, die sehr scharf an der Grenze des Erlaubten und Haftbaren agieren. Kontrolle passiert grundsätzlich erst durch Gerichte im Rahmen von Klagen. Dadurch entstehen auch immer wieder Schäden für Private, die entsprechend das Vertrauen in gewissen, speziell angehängten, Gruppen der Gesellschaft reduzieren und diese Abgehängten empfänglich für Propaganda macht.

Ansonsten wird auch im privaten AI genutzt, ohne sich viel Gedanken zu machen, Bezahlen mit privaten Daten ist normal, diese werden auch sonst sehr freizügig auf Social Media geshared. Die Nutzung dieser Daten durch AI-Betreiber ist normal und nicht verpönt.

Auch LEAs bekommen sehr schnell gute neue Werkzeuge, grundsätzlich können diese aus EU-weiter heimischer Produktion erstanden werden, entsprechend besitzt LEAs auch einiges an Kontrolle über die Produkte, allerdings fehlen generelle Regularien zu Sicherheit und Datenschutz, diese müssen direkt von den LEAs für ihre eigenen Tools verlangt und durchgesetzt werden. Dadurch kommt es immer wieder zu Skandalen vor Gericht oder bei der Nutzung unsicherer Software. In Hinblick auf den Talente-Pool ist dieser sehr groß und Talente können auch im Polizeidienst gut gehalten werden, allerdings sterben viele Low-level-Jobs im Technologiebereich, bspw. Junior Developer aus. Dies wird mittel- und langfristig Probleme zeitigen, sollte es nicht gelingen, auch die senioren Positionen entweder durch AI zu besetzten oder anderweitig zu trainieren.

Generelle Effekte auf LEAs

Folgen eine Skizzierung der möglichen extrapolierten Effekte auf Law Enforcement Agencies, basierend auf der Gestalt des vorliegenden Szenarios, zusammen mit kurzen Anmerkungen und Beschreibungen, wo dies für notwendig erachtet wurde. Dabei sind viele Effekte sehr ähnlich denen aus dem ersten Szenario, teilweise nur etwas extremer ausgeprägt, wir bitten daher, auch in diesem Szenario nachzuschlagen und kompatible Effekte zu übernehmen, wir haben hier aus Gründen der Übersichtlichkeit einige Aspekte kurzgehalten.

- Sehr guter Talente-Pool: Siehe auch Szenario 1, allerdings fehlt den Talenten oftmals das Bewusstsein für die Gefahren, die unsachgemäßer Einsatz ihres Wissens erzeugen können.
- Günstige Werkzeuge: Ebenfalls wie in Szenario 1, allerdings ist die Qualität der Werkzeuge deutlich verschieden und tatsächlich oft unzureichend.
- AI-Einsatz wird alltäglich, allerdings ist auch hierbei der Einsatz nicht immer gerechtfertigt und wird auch ohne viel Rücksicht auf andere Rahmenparameter durchgeführt. Oftmals wird es den Einsetzenden nicht bewusst sein, dass sie gerade Ai nutzen.
- Viele Fehler durch neue Werkzeuge oder unsachgemäße Handhabung: Ein Nachteil der sofortigen Verfügbarkeit immer neuer Werkzeuge mit dem Klima des Szenarios ist die Anfälligkeit, fehlerhafte Werkzeuge zu nutzen und u.U. sogar das Entstehen von Fehlern in der Ermittlung bewusst zu tolerieren, d.h. nach der Perspektive, dass Performance und Features wichtiger sind als unbedingt Korrektheit auch das Risiko für Fehler bewusst in Kauf zu nehmen um weitere Features zu bekommen und/oder effizienter zu werden. Da wir annehmen, dass sich die Gesellschaft in diesem Szenario nicht von einem Rechtsstaat wegentwickeln wird, kann dies zu erheblichen Aufwänden führen (siehe auch unten) und auch negative Effekte auf die Wahrnehmung der AI-Nutzung durch LEAs haben.
- Sehr viel Nachbearbeitung durch Intransparenzen: In Zusammenhang mit dem vorangegangenen Effekt kann davon ausgegangen werden, dass es öfter zu notwendigen Nacharbeiten auf Seite der LEAs kommen wird, sollte sich ein Tool als fehlerhaft herausstellen, d.h. Fälle, in denen das Tool zum Einsatz kam werden nachbearbeitet und u.U. wiederaufgerollt werden müssen, was erhebliche Kosten verursachen kann. Dies kann sich auch grundlegend auf die Art, wie Verfahren und/oder Gutachten in Zukunft gehandhabt werden auswirken.

- Viele Low-Level-Angriffe verfangen nicht, da auch normale Bürger typischerweise in der Lage sind, diese zu erkenne, speziell im Bereich der Fake News und Grassrooting, sowie spezifisch im Bereich des Phishings mit AI.

- Neue große Angriffsflächen in LEAs: Die Angriffsfläche der LEAs könnte sich in diesem Szenario noch weiter vergrößern: Erstens, gehen wir in diesem Szenario davon aus, dass Tools nicht immer sehr sorgfältig entwickelt wurden, sogar wenn diese für den polizeilichen Bereich gedacht wurden. Dabei gehen wir in diesem Szenario davon aus, dass Performance, Features gegenüber Safet und Security bevorzugt behandelt werden. Zweitens gehen wir davon aus, dass im Klima dieses Szenarios auch Werkzeuge durch Mitglieder von LEAs für ihre Arbeit verwendet werden, die niemals für den Einsatz in einer derart heiklen Umgebung gedacht waren und entsprechend auch nicht nach internen Sicherheits-Maßgaben der LEAs entwickelt wurden. Drittens gehen wir davon aus, dass auch Mitglieder von LEAs sich im Privatleben wie normale Bürger*Innen verhalten, speziell was das Teilen von Informationen im Internet und mit AI-Systemen angeht. Dabei eröffnen sich natürlich einige Angriffsmöglichkeiten auf diese Mitarbeiter*Innen basierend auf privaten Informationen oder in einem privaten Setting.

- Verbesserte Aufklärung (spez. OSINT): Wie in Szenario 1, allerdings mit noch mehr Unterstützung durch Tools auf Seite des Law Enforcements, dafür einer reduzierten Genauigkeit und Zuverlässigkeit.

- Datenquellen sind von unterschiedlicher Zuverlässigkeit: Die Qualität der Datenquellen nehmen wir in diesem Szenario genauso an wie die Qualität der Tools und Modelle: Wechselhaft, je nachdem, woher die Daten stammen und welche Qualitätskriterien an sie angelegt wurden. Wir gehen in diesem Szenario davon aus, dass LEAs auch Daten schlechterer oder unbekannter/unkontrollierbarer Qualität nutzen, da dies dem Geiste der Gesellschaft in diesem Szenario entspricht und wir davon ausgehen, dass eine Vielzahl nützlicher Quellen zweifelhafter Qualität zur Verfügung stehen werden, die weitere Features und Performancesteigerungen ermöglichen, allerdings auf Kosten der Zuverlässigkeit.

- Nutzung von AI in der Bevölkerung durch LEAs zwar grundsätzlich akzeptiert, aber je nach Anzahl und Intensität der Fehler bei der Nutzung kann dies auch ins Gegenteil umschlagen, da im Gegensatz zu Szenario 1 die starke Kontrolle reduziert wurde und eine hohe Fehlertoleranz herrscht.

- Kriminelle analysieren LEA-Verhalten mit AI: Wie in Szenario 1 nutzen auch kriminelle AI zur Analyse von LEA-Verhalten, zusätzlich könnten sie dies in Kombination mit Angriffen auf eines der mannigfaltigen genutzten Tools im Law Enforcement kombinieren.

Ausgewählte neue Anwendungen für LEAs

Folgend geben wir einen kurzen Überblick darüber, welche neuen Anwendungen wir in dem gegenständlichen Szenario als sehr wahrscheinlich erachten. Grundsätzlich ist in diesem Szenario dn Möglichkeiten der Anwendung keine Grenzen gesetzt, alle bisher genannten Anwendungen wären daher möglich, zusätzlich aber auch jene, die derzeit entweder konkret verboten sind (AI-Act, GDPR), oder aber derart heikel sind, dass sie in der derzeitigen Realität praktisch unmöglich durch LEAs genutzt werden können, bspw:

- Predictive Policing
- Crime Forecasting

Lücken und resultierende Herausforderungen

Folgend geben wir einen kurzen Überblick darüber, welche neuen Lücken, Gefahren und Herausforderungen wir in dem gegenständlichen Szenario als sehr wahrscheinlich erachten. Eine Beschreibung findet sich im Anschluss an die Szenarienbeschreibung.

- Austrocknung des Talente-Pools
- Security-Testing von AI ressourcentechnisch komplex
- Sehr große Datenmengen mit viel Noise
- Viele aktive Cyber-Kriminelle
- Problematische neue Angriffe via shared Devices
- Mangelhafte Kontrolle von Einzelnen in LEAs
- Unbeabsichtigte Informationsweitergabe
- Absicherung gegen Re-Extraktion von Trainingsdaten
- Weitergabe strategischer Meta-Informationen durch gemeinsame Nutzung von neuen Anwendungen
- Proaktive Erkennung privater Informationen schwierig
- Absicherung gegen Weitergabe der gesharten Daten
- Absicherung gegen Backdoors in pre-trained models
- Private Systeme werden u.U. Angriffsziel
- Angriffe durch präparierte gesharte Geräte
- Attack- und Backdoor-Attribution schwierig
- AI-generiertes dezentrales data flooding erkennen
- Fake-News und Information Warfare zielt auf Abgehängte ab
- Angriffsoberfläche im Bereich der LEAs wächst
- Blind Spots durch Analyse der LEA-Werkzeuge
- AI-generiertes dezentrales data flooding erkennen

Szenario 5 – „Geopolitische Instabilität“

Ausgewählte Faktorausprägungen:

Dieses Szenario wurde erst am Ende unserer Analyse vor dem Hintergrund der derzeitigen politischen Umwälzungen entworfen und zielt darauf ab, dass speziell die am Ende der Interview-Periode interviewten Expert*Innen deutliche Sorgen vor politischer Instabilität äußerten. Tatsächlich war dies eine Schwäche in den Interviews, dass sie zu einer Zeit politischer Veränderungen durchgeführt wurden, d.h. während dieser Faktor am Anfang der Interview-Phase kein großes Thema darstellte, war dies in der letzten Phase ein bestimmender Aspekt. Dieser eine Faktor, die geopolitische Instabilität wurde von uns in den Interviews am Ende auch als sehr dominant wahrgenommen, viel mehr noch als die Deregulierung in Szenario 4 und viele Faktoren waren nicht mehr frei wählbar sondern nur in spezifischen Faktorausprägungen. Zusätzlich ist diese Faktorausprägung sehr breit und unspezifisch, daher ist das entworfene Szenario noch weitaus freier entworfen als die vorherigen, wo wir zumindestens basierend auf unserer Ausbildung Effekte, Anwendungen und Herausforderungen skizzieren konnten. Auch fehlt die Erfahrung, wie Europa mit so einer Situation tatsächlich umgeht und welche Paradigmenwechsel tatsächlich stattfinden (können), wobei dies speziell auch von Art und Intensität der Instabilität abhängt. All dies macht natürlich dieses Szenario zum Spekulativsten, allerdings wollten wir aus Gründen der Vollständigkeit und der

Aktualität vieler Themen nicht darauf verzichten. Wir haben allerdings aus dem Faktor, entgegen der rein formalen Analyse, basierend auf den Gesprächen Sub-Faktorausprägungen extrahiert, dieses Szenario bricht daher auch methodisch mit den vorangegangenen.

- Geopolitische Instabilität als treibenster Faktor dominiert
    - Regulierung von AI bezogen auf Anwendungsgebiete und Notwendigkeiten
    - Abwägung Qualität / Performance
    - Aufbau einer entsprechend ausgerichteten Industrie
    - Jobs in Verwaltung und Low-level-Coding aus Resilienzgründen erhalten
    - LEAs bekommen benötigte Werkzeuge schnell, Sicherheit wichtig
    - Einsatz von AI bei LEAs wird stark ge-monitored
    - LEAS haben Backdoors in den wichtigsten AI-Anwendungen
- Kosten von AI durch Algorithmen und Hardware für sehr viele Anwendungen ausreichend gering
- Qualität ist für die meisten Branchen ausreichend erzielbar
- Auswirkungen auf den Arbeitsmarkt Nullsummenspiel, aber einige Abgehängte
- Privacy wird auch für Private wichtiger
- Kaum Sharing-economy im digitalen Beriech
- AI auch im privaten Bereich sehr populär
- AI-Systeme müssen zugelassen und kontrollierbar sein
- Abschottung der Bürger*Innen gegen Fremdsysteme
- Sehr starker Fokus auf Security
- Datenqualität ist ein wesentlicher Faktor

Szenariobeschreibung

Dieses Szenario ist sicher in vielen Aspekten deutlich unterschiedlich von anderen und war vor allem in der Konsistenzanalyse spannend, da ein überbordend wichtiger Faktor alle andern dominierte, nämlich der der geopolitischen Stabilität in seiner Faktorausprägung der geopolitischen Instabilität. Dabei geht es nicht um leichte politische Irritationen, sondern um Instabilität am Rande eines Krieges, bzw. mit starken Animositäten zu zentralen Machtblöcken, u.U . auch ehemaligen Verbündeten. Dadurch befindet sich die Gesellschaft zwar noch nicht in einem echten, formal ausgerufenen Kriegsrecht, viele Gesetze und Verordnungen unterwerfen sich jedoch bereits ähnlichen Sichtweisen und fokussieren sich auf das Ziel der Verteidigung. Dies führt dazu, dass bspw. Regularien nicht mehr anwendungsneutral und „fair“ in Hinblick auf Rahmenparameter wie Betreiber, Anwendungsgebiet und Standort beurteilt werden, sondern vor dem Hintergrund der Instabilitäten nicht mehr neutral, sondern rein utilaristisch argumentiert und reguliert wird, d.h. was in einem Anwendungsgebiet und/oder einer heimischen Firma erlaubt wird, muss demnach nicht so für andere Gebiete/Anbieter gelten.

In diesem Szenario wird eine sehr leistungsstarke AI-Industrie aufgebaut, die sich nur mehr bedingt dem Markt unterwerfen muss, sondern hochsubventioniert ist, wobei ein spezielles Augenmerk auf der Kontrolle über die entstehenden Systeme gelegt wird – sowohl in Bezug auf Security-Anforderungen als auch in Bezug auf tatsächliche direkte Kontrolle bspw. durch Backdoors für LEAs. Umgekehrt bekommen LEAs zwar Werkzeuge sehr schnell, Security und Reduktion der Angriffsoberfläche sind aber wesentliche Entwicklungsaspekte, zusätzlich wird der Einsatz der Werkzeuge stark kontrolliert, um gegen potenzielle Feinde von Innen gewappnet zu sein. Dies wiederum führt zu einem gewissen Klima der Kontrolle und des Misstrauens innerhalb der Organisationen.

Kosten für AI sind in diesem Szenario für viele Anwendungen ausreichend gering, nicht nur durch den Aufbau einer starken Industrie bedingt, sondern auch, weil vom Staat oder der Europäischen Union als kritisch eingestufte Anwendungen nicht nach den Gesetzen des Marktes spielen müssen, sondern großzügige Förderungen erhalten. Dies bedeutet auch, dass die Qualität der als wichtig erachteten Anwendungen üblicherweise ausreichend hoch ist. Zusätzlich müssen AI-Systeme zentral erfasst und zugelassen werden, was die Umsetzung dieser Regularien, aber auch von Qualitätskontrollen deutlich erleichtert, allerdings natürlich auch eine Vielzahl neuer, potenzieller Delikte zeitigt, spezilel bei der Nutzung von Online-Services durch normale Bürger*Innen.

In einem gewissen Kontrast zu der starken Nutzung von AI durch LEAs und auch bedingt durch die starke Kontrolle kommt es zu einer bewusst herbeigeführten Abschottung von Bürger*Innen gegenüber Fremdsystemen, d.h. Systemen, die nicht der Kontrolle der heimischen Behörden unterliegen. Dies kann zum einen bewusst und explizit verordnet werden, bspw. In Hinblick auf Social Media Plattformen, oder aber implizit durch Erschwerung der Nutzung, bzw. Forcierung alternativer, heimischer Angebote, geschehen. Diese Art der Forcierung heimischer Systeme sehen wir auch jetzt schon in anderen Staaten außerhalb der Europäischen Union.

Ein weiterer wesentlicher Unterschied zu den vorangegangenen Szenarien ist das Verhalten von Privatpersonen in diesem Szenario: Die politischen Spannungen haben den Normhaushalt erreicht und die Gefährdungslage ist auch dort bewusst, daher wird Privacy auch für Private wichtig, das Teilen von Ressourcen im Rahmen einer Sharing-Economy leidet unter diesen Ansätzen, da es nicht vereinbar mit der gestiegenen Verantwortung und den Kontrollanforderungen ist.

Letztlich ändern sich in diesem Umfeld auch die Aufgaben der LEAs drastisch, es entstehen neue Delikte und die Strafmaße/Relevanz aller Delikte steigt oder sinkt.

Anmerkung: In diesem Szenario wurde auch eine gewisse Limitierung des originalen Ansatzes von Geschka, aber auch unseres originalen Ansatzes in der modifizierten Form offenbar. Oftmals wurden Faktorausprägungen als inkonsistent angenommen, weil sie in unserer gegenwärtigen Gesellschaft als Inkonsistent angesehen werden würden, bspw. Der Trend zu extrem starker AI-Nutzung bei gleichzeitiger Erstarkung des Privacy-Gedankens. Ein weiteres Beispiel wäre eine sehr schnelle Verfügbarkeit von Werkzeugen für LEAs bei gleichzeitig sehr starker Einschränkung und Kontrolle der Industrie, oder aber der Aufbau einer stark subventionierten Industrie, die die Nutzung von Fremdsystemen zu reduzieren trachtet. Dies ist auf den einen dominierenden Faktor zurückzuführen, der quasi die implizit angenommenen Spielregeln einer Gesellschaft verändert. Wir werden diesen Aspekt noch in einer späteren Version der Modellierung aufnehmen, derzeit vielleicht als Meta-Faktor oder Meta-Bedingung geplant. Da dieses Szenario erst sehr am Ende der Forschungsperiode aufkam, konnte dies leider nicht ausreichend formalisiert werden.

Generelle Effekte auf LEAs

Im Folgenden haben wir skizziert, welche generellen Effekte auf LEAs in unserem Szenario wir als sinnvoll und wahrscheinlich erachtet haben, allerdings muss hier noch stärker als in den vorangegangenen Szenarien darauf hingewiesen werden, dass Extrapolation der Zukunft keine exakte Wissenschaft ist, sondern viel Meinung und „educated guessing“ einfließt – in diesem Fall noch deutlich mehr, da keiner der Beteiligten (zum Glück) ein derartiges Szenario auch nur im Ansatz erlebt hat.

- Aufgaben ändern sich drastisch: LEAs bekommen neue Aufgaben und dienen auch als Verteidigung gegen echte oder wahrgenommene Feinde von Innen. Es muss davon ausgegangen werden, dass wesentlich mehr präventive Strafen verteilt werden, gleichzeitig aber auch die Überwachung im Inneren ein wesentlich größeres Thema wird. Zusätzlich ist in unserem Szenario keine wesentlich größere Härte gegenüber Bürgern avisiert, da dies nur die Unzufriedenheit stärken würde, vielmehr geht man von einer pluralistischen Gesellschaft hin zu einer Reduktion der Aussagefreiheit in Bezug auf gewisse geopolitische Themen. Wir nehmen in diesem Szenario noch an, dass dies hauptsächlich über gesellschaftlichen Druck geschieht, nicht über polizeistaatliche Maßnahmen, derartige Aufgaben für LEAs sehen wir in diesem Szenario nicht vor. Wir gehen auch immer noch von einem Rechtsstaat aus, allerdings mit reduzierter Toleranz gegenüber gewissen Meinungen und einer Einschränkung gewisser Freiheiten.
- Information Warfare als großes Thema: Als Spezialisierung des vorangegangenen Effekts sehen wir die Aufgaben der LEAs zwar weiterhin im zivilen Bereich und nicht in der militärischen Abwehr, allerdings werden Aspekte wie Terrorbekämpfung wesentlich mehr Raum einnehmen und auch als wichtiger erachtet werden. Dabei wird in unserem Szenario auch die Verbreitung von Propaganda, aber auch die subversive Meinungsbildung und Meinungsbeeinflussung via Social Media und dergleichen potenziell als Information Warfare eingestuft werden. Gleichzeitig werden andere, nicht der Polizei unterstellte staatliche Akteure, eigene Propaganda nach Innen spielen.
- Zensur- und Filteraktivitäten: Diese neuen Aufgaben für LEAs werden wesentliche Ressourcen binden, aber auch von der Automatisierung mittels AI wesentlich profitieren. Dabei werden ethische und rechtliche Fragestellungen berührt werden, die für die LEAs bisher Neuland waren und viele Verfahren bis zu höchstrichterlichen Entscheidungen hin erfordern. Die wird sich auch auf die Arbeit und Performance der entsprechenden Stellen auswirken, da eine gewisse Rechts- und Verfahrensunsicherheit bestehen bleiben wird. Starke Kontrollen der LEAs verhindern aber in unserm Szenario Willkür, ein starker Rechtsstaat bleibt erhalten.
- Mittelmäßiger Talente-Pool für LEAs, viele gute Leute gehen direkt in die Industrie, für die es sehr viel Arbeit gibt, speziell da auch fremdbestimmte Plattformen abgelöst werden sollen, gleichzeitig durch die hohen Subventionierungen wesentlich bessere Gehälter möglich sind. Gleichzeitig könnten viele Expert*Innen dem Klima der geopolitischen Instabilität dadurch zu entkommen versuchen, dass sie ihren Lebensmittelpunkt woanders hin verlegen. Ob dies eine sinnvolle und durchführbare Option ist, ist direkt abhängig von der genauen Gestalt der Instabilität, sowie der davon betroffenen Staaten und kann von einer kaum wahrnehmbaren bis hin zu einer starken Bewegung reichen. In unserem Szenario haben wir uns für ersteres entscheiden, vor dem Hintergrund, dass wir immer noch von einem wohlhabenden Rechtsstaat mit großen Freiheiten ausgehen und gleichzeitig keine immenente echte Kriegsgefahr angenommen wird.
- Vertrauenswürdige und günstige Werkzeuge stehen LEAs zur Verfügung, wobei Security ein wesentlicher Beschaffungsaspekt ist. Der Preis ist nicht nur durch eine etablierte Industrie erzielbar, sondern wird in Randthemen, bei denen der Markt nicht für eine kostendeckende Entwicklung sorgen kann, durch Subventionen künstlich niedrig gehalten.
- AI-Einsatz wird alltäglich, viele Aufgaben von LEAs werden automatisiert und diese Automatisierung wird auch von den meisten Bürger*Innen als notwendig erachtet.
- Kosten spielen weniger eine Rolle, wie schon im Szenario selbst beschrieben löst die Krise einen Boom in der AI-Industrie aus, spezielle Anwendungen werden durch Subventionen entwickelbar gemacht.

- Militarisierung der LEAs: Auch wenn in diesem Szenario der Rechtsstaat erhalten bleibt, so werden den LEAs aus Kostengründen viele Tools zur Verfügung gestellt, die einen Dual Use haben, d.h. die auch mit militärischen Anwendungsfällen als Hintergedanke implementiert wurden Diese Möglichkeiten müssen stark kontrolliert eingesetzt werden, um eine Militarisierung der LEAs zu verhindern.
- Nutzung von AI durch LEAs ist in der Bevölkerung akzeptiert, speziell da dies an das Schutzbedürfnis der Bevölkerung appelliert, d.h. das Verhalten der Behörden wird von der eigenen Bevölkerung als deutlich weniger kritisch gesehen, bzw. ist die Bevölkerung dazu bereit in einem gewissen Ausmaß und für eine gewisse Zeitdauer den Behörden mehr Rechte einzuräumen, als dies derzeit der Fall ist. Dabei gehen wir aber davon aus, dass starke Kontrollen gegenüber Machtmissbrauch ein wesentlicher Faktor für den Erhalt dieser Akzeptanz ist.
- Reduktion auf abgesicherte Datenquellen und rein europäische Tool/Daten-Chains: Auch die Datenquellen werden Audits unterzogen werden, die Datengenerierung in europäische Hände gelegt, bspw. durch eigene Satelliteninfrastrukturen oder der Fertigung eigener Netzwerksensoriken. Inwieweit es möglich sein wird, selbst in diesem Szenario ein rein europäisches digitales Ökosystem aufzubauen bleibt diskutierbar, allerdings werden Fremdkomponenten rigorosen Sicherheitsanalysen unterzogen werden.
- Große Härte auch gegen kleine Cyber-Kriminelle: Cyber-Crime, speziell Phishing und Attacken gegen Firmen, werden vermehrt in einem Umfeld potenzieller fremdstaatlicher Einmischung angesiedelt, speziell im Fall von Ransomware und dergleichen, da diese auch schon jetzt von gewissen Staaten stark unterstützt, oder sogar zu Devisengewinnung staatlich genutzt werden. Daher wird auch in Fällen der rein privaten Bereicherung sehr schnell in Richtung Terrorismus und/oder fremdstaatliche Einmischung ermittelt werden, was die Ermittlungen entsprechend verzögern wird. Zusätzlich könnte dies zu einer weit großzügigeren Anwendung von Untersuchungshaft führen.

Neue Anwendungen für LEAs

Folgend geben wir einen kurzen Überblick darüber, welche neuen Anwendungen wir in dem gegenständlichen Szenario als sehr wahrscheinlich erachten. Dabei sind dies nur Beispiele, vieles wird von Art und Intensität der geopolitischen Instabilität abhängen.

- Netzkontrolle und Network Monitoring mit AI
- Kontrollierbarmachung von AI, Monitoring von AI
- Erkennung und Abschottung von Netzwerken
- Information Warfare

Lücken und resultierende Herausforderungen

Auch hierbei kommt es natürlich sehr stak auf die tatsächliche Intensität und vor allem auch Dauer der geopolitischen Instabilität an, dennoch möchten wir ein paar wesentliche Gaps herausstellen, die so in den anderen Szenarien nicht vorkommen.

- Ethische Probleme, überbordende Strenge
- Arms Race mit staatlichen Akteuren
- Hoher Testing-Aufwand für neue AI-Anwendungen
- Abschottung/Kontrolle sehr kompliziert
- Zensur- und Informationsfilter komplex

- Potenziell mangelhafte Kontrolle Einzelner in den LEAs
- Insider-Angriffe
- Reduktion im Wissensaustausch
- Reduzierte Möglichkeiten durch Einschränkung auf eigene Tools und Modelle

Beschaffungsleitfaden

Einleitung

Die fortschreitende Digitalisierung und die Integration von künstlicher Intelligenz (KI) in nahezu alle Geschäftsbereiche stellen Unternehmen, Behörden und Betreiber kritischer Infrastrukturen vor neue Herausforderungen. KI-Systeme bieten enorme Chancen – von der Optimierung von Prozessen über die Automatisierung komplexer Aufgaben bis hin zur Unterstützung bei sicherheitsrelevanten Entscheidungen. Insbesondere bei High-Risk-Anwendungen, etwa in der öffentlichen Sicherheit, der Gesundheitsversorgung oder der kritischen Infrastruktur, ist der zuverlässige und sichere Betrieb von KI-Systemen von zentraler Bedeutung.

Da viele Organisationen nicht über die internen Ressourcen oder das nötige Know-how zur Eigenentwicklung von KI-Lösungen verfügen, gewinnt der externe Erwerb zunehmend an Bedeutung. Der vorliegende Beschaffungsleitfaden für sichere KI zielt darauf ab, einen systematischen und transparenten Prozess für die Auswahl und Beschaffung externer KI-Lösungen zu etablieren. Dabei wird auch unser eigens entwickelter Make, Buy or Transfer Guide integriert, der als strategische Entscheidungshilfe dient und unterschiedliche Beschaffungsoptionen abwägt.

Grundlagen und Rahmenbedingungen

Der Leitfaden richtet sich an Entscheidungsträger in Unternehmen und Behörden, die KI-Systeme in sicherheitskritischen Infrastrukturen einsetzen möchten. Ziel ist es, den gesamten Beschaffungsprozess strukturiert darzustellen und so sicherzustellen, dass alle relevanten Anforderungen – organisatorisch, technisch und rechtlich – berücksichtigt werden. Der Fokus liegt dabei auf High-Risk-Anwendungen, bei denen Fehlentscheidungen oder Sicherheitslücken gravierende Folgen haben können.

Rechtliche und regulatorische Grundlagen

Die Beschaffung und der Einsatz von KI-Systemen unterliegen zahlreichen rechtlichen Rahmenbedingungen. Neben der Datenschutz-Grundverordnung (DSGVO) sind insbesondere die im AI-Act, Data Act und anderen relevanten Regularien festgelegten Anforderungen zu beachten. In unseren Dokumenten wurden wichtige Aspekte dieser Regularien herausgearbeitet. Zu den zentralen rechtlichen Anforderungen zählen:

- Datenschutz und Datensicherheit: Einhaltung von Datenschutzrichtlinien und Implementierung technischer Maßnahmen (z. B. Verschlüsselung, Zugriffskontrolle).
- Haftungsregelungen: Klare Definition von Haftungsfragen im Falle von Fehlfunktionen oder Sicherheitsvorfällen.
- Compliance und Zertifizierung: Regelmäßige Überprüfung der Einhaltung gesetzlicher Vorgaben und gegebenenfalls externe Zertifizierungen (z. B. NIST AI Risk Management Framework, ISO/IEC 23894:2023 – AI Risk Management Guidance, ISO/IEC 27001).

Organisatorische und technische Voraussetzungen

Neben rechtlichen Aspekten spielen organisatorische Strukturen und technische Standards eine zentrale Rolle. Der Aufbau eines interdisziplinären Teams, bestehend aus IT-Experten, Juristen, Compliance-Beauftragten und Fachbereichsvertretern, ist unerlässlich. Organisatorisch sollten transparente Prozesse und regelmäßige Schulungen implementiert werden, um alle Beteiligten auf den aktuellen Stand zu bringen. Technisch ist es wichtig, dass die zu beschaffende Systeme den Anforderungen an Sicherheit, Skalierbarkeit und Integrationsfähigkeit genügen.

Ermittlung der Anforderungen

Analyse der internen Bedürfnisse

Der erste Schritt im Beschaffungsprozess besteht darin, eine detaillierte Anforderungsanalyse durchzuführen. Hierzu gehören:

- **Zieldefinition:** Welche konkreten Probleme sollen durch den Einsatz der KI gelöst werden? Welche operativen oder strategischen Ziele stehen im Vordergrund?
- **Anwendungsfälle:** Beschreibung der spezifischen Use Cases, z. B. Überwachung kritischer Infrastrukturen, Entscheidungsunterstützung in sicherheitsrelevanten Bereichen oder automatisierte Analysen großer Datenmengen.
- **Sicherheits- und Qualitätsanforderungen:** Welche Sicherheitsstandards und Qualitätskriterien müssen erfüllt sein, um den hohen Anforderungen kritischer Systeme gerecht zu werden?
- **Rechtliche Vorgaben:** Welche gesetzlichen und regulatorischen Rahmenbedingungen sind zu beachten? Dies umfasst sowohl nationale als auch EU-weite Vorgaben.

Stakeholder-Analyse

Es ist entscheidend, alle relevanten Stakeholder in den Beschaffungsprozess einzubinden. Hierzu zählen:

- **Interne Abteilungen:** IT, Sicherheit, Recht, Controlling und operative Fachbereiche.
- **Externe Experten:** Externe Berater, Technologieanbieter und gegebenenfalls wissenschaftliche Partner.
- **Regulierungsbehörden:** Kontakt zu Aufsichtsbehörden, um frühzeitig regulatorische Anforderungen zu klären und zukünftige Entwicklungen zu antizipieren.

4 Marktanalyse und Anbieterrecherche

Identifikation potenzieller Anbieter

Nach der Festlegung der Anforderungen folgt die Marktanalyse. Dabei sollten potenzielle Anbieter von KI-Lösungen systematisch identifiziert und bewertet werden. Wichtige Kriterien sind:

- **Technologische Kompetenz:** Wie innovativ und robust ist die angebotene Lösung?
- **Erfahrungen und Referenzen:** Welche Erfahrungen haben die Anbieter bereits im Bereich kritischer Infrastrukturen gesammelt?
- **Compliance und Sicherheitsstandards:** Erfüllt die Lösung die geltenden rechtlichen Vorgaben und Sicherheitsanforderungen?
- **Integrationsfähigkeit:** Wie gut lässt sich die Lösung in bestehende IT-Landschaften integrieren?
- **Kosten und Wirtschaftlichkeit:** Welche Investitions- und Betriebskosten fallen an?

Durchführung einer Marktanalyse

Die Marktanalyse sollte quantitative und qualitative Elemente umfassen. Neben der Auswertung von Marktdaten und Benchmarks empfiehlt es sich, Anbieter zu kontaktieren und Referenzprojekte anzusehen. Dies ermöglicht ein umfassendes Bild über die Leistungsfähigkeit der jeweiligen Lösungen und liefert wichtige Hinweise zur praktischen Umsetzbarkeit.

5 Strategische Entscheidungsfindung: Make, Buy or Transfer Guide

Ein zentraler Bestandteil des Beschaffungsprozesses ist die strategische Entscheidung, ob eine Lösung intern entwickelt („Make“), extern beschafft („Buy“) oder von einem bestehenden System übernommen bzw. angepasst werden soll („Transfer“). Unser integrierter Make, Buy or Transfer Guide dient als Entscheidungshilfe.

Make-Buy-or Transfer Guide für KI

Einleitung

Die zunehmende Digitalisierung und Automatisierung kritischer Infrastrukturen (KRITIS) stellt Entscheidungsträger vor die Herausforderung, ob technologische Lösungen intern entwickelt (Make), extern beschafft (Buy) oder von Partnern bzw. anderen Institutionen übernommen (Transfer) werden sollen. Jede dieser Optionen bringt spezifische Vor- und Nachteile sowie rechtliche Verpflichtungen mit sich, insbesondere im Kontext des AI-Act und der regulatorischen Anforderungen der Europäischen Union. Ein fundierter Entscheidungsprozess kann nicht nur Kosten und Risiken minimieren, sondern auch die Sicherheit und Resilienz kritischer Infrastrukturen langfristig gewährleisten.

Dieser Leitfaden soll eine strukturierte Entscheidungsgrundlage bieten, indem er die wichtigsten Aspekte der drei Optionen beleuchtet und klare Empfehlungen zur Auswahl der passenden Strategie gibt. Darüber hinaus werden Beispiele aus der Praxis sowie mögliche Anwendungen in Ungarn betrachtet, um eine übertragbare und praxisnahe Orientierung zu ermöglichen.

Make-Buy-or Transfer Guide für KI

Die folgende Entscheidungsstruktur hilft, die bestmögliche Lösung für eine bestimmte Organisation oder Infrastruktur zu identifizieren:

1. **Technische Kapazitäten:** Verfügt die Organisation über eigene F&E-Ressourcen und IT-Fachkräfte?
2. **Regulatorische Anforderungen:** Gibt es gesetzliche Vorgaben zu Datenschutz, Datensouveränität oder Sicherheitsstandards?
3. **Anpassungsbedarf:** Benötigt die Lösung spezifische Anpassungen oder individuelle Konfigurationen?
4. **Marktverfügbarkeit:** Gibt es etablierte Lösungen am Markt, die die Anforderungen erfüllen?
5. **Kosten-Nutzen-Analyse:** Ist die langfristige Wartung und Weiterentwicklung wirtschaftlich tragfähig?

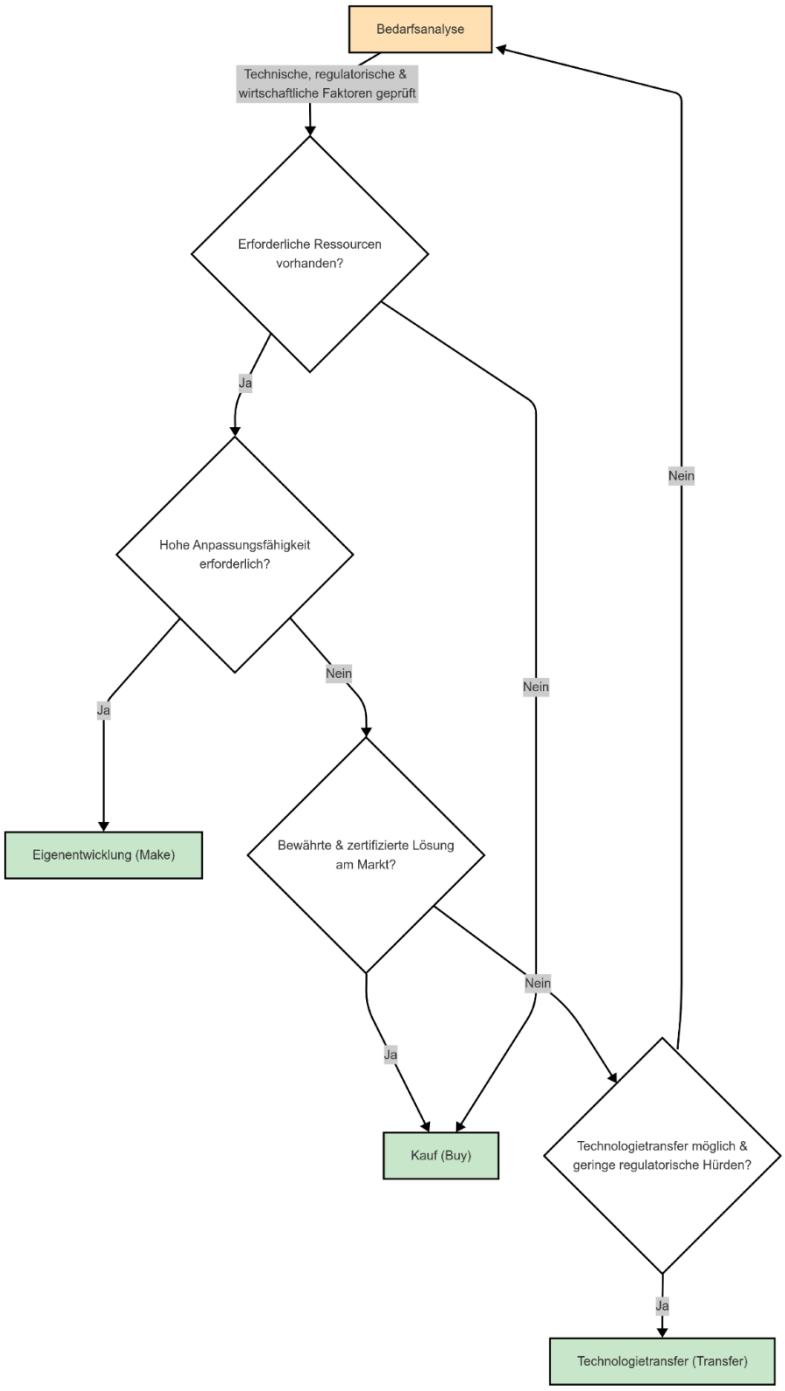


Abbildung 41: Ablauf

Falls die Organisation starke interne Kapazitäten hat und hohe Anpassbarkeit erfordert, kann eine Eigenentwicklung sinnvoll sein. Ist hingegen eine schnelle Umsetzung erforderlich und eine zuverlässige Lösung am Markt vorhanden, ist der Kauf vorzuziehen. Falls eine Technologie bereits existiert und durch Kooperation oder staatliche Unterstützung erworben werden kann, könnte eine Transferlösung die beste Wahl sein.

Der Entscheidungsbaum ist folgendermaßen strukturiert:

- Falls eine Organisation über hohe IT-Kompetenz, regulatorische Kontrolle und spezifische Anpassungsbedarfe verfügt, ist **Make** die bevorzugte Option.

- Wenn eine bewährte Lösung am Markt verfügbar ist, die alle Anforderungen erfüllt, sollte **Buy** in Betracht gezogen werden.
- Falls eine bereits existierende, erprobte Technologie von einem Partner übernommen werden kann und regulatorische Hürden gering sind, bietet sich **Transfer** als geeignete Lösung an.

Option 1: Make

Die Option 1: Make oder Eigenentwicklung einer KI-Lösung bedeutet, dass eine Organisation die Software intern entwickelt, anpasst und betreibt. Dies bietet maximale Kontrolle über Sicherheit, Datenschutz und Anpassungsfähigkeit an spezifische Anforderungen. Allerdings erfordert diese Option hohe Investitionen in Entwicklung, Fachkräfte und Wartung. Sie ist besonders sinnvoll, wenn bestehende Lösungen nicht die erforderlichen Sicherheits- oder Compliance-Standards erfüllen.

Vorteile:

- Volle Kontrolle über Technologie, Daten und Anpassungen
- Keine Abhängigkeit von externen Anbietern
- Möglichkeit, sicherheitskritische Anforderungen maßgeschneidert zu erfüllen
- Langfristig potenziell kosteneffizienter bei nachhaltiger Nutzung

Nachteile:

- Hoher initialer Entwicklungsaufwand
- Erfordert spezialisierte IT- und Sicherheitsexperten
- Wartung und Weiterentwicklung sind langfristige Verpflichtungen
- Regulatorische Verantwortung für Datenschutz, Sicherheit und Transparenz

Rechtliche Verantwortlichkeiten:

- Einhaltung des AI-Act, insbesondere Art. 9 (Risikomanagement) und Art. 27 (Grundrechte-Folgenabschätzung)
- Sicherstellung der DSGVO-Konformität
- Dokumentations- und Auditpflichten für behördliche Kontrollen
- Haftung für Fehlentscheidungen des KI-Systems

Beispiel:

Ein nationaler Netzbetreiber entscheidet sich, ein eigenes KI-basiertes Frühwarnsystem für Cyberangriffe zu entwickeln. Dies ermöglicht maximale Kontrolle über Algorithmen und Datenverarbeitung, erfordert jedoch eine erhebliche Investition in Forschung und Entwicklung sowie laufende Compliance-Prüfungen.

Option 2: Buy

Die Option 2: Buy ist Fremdbeschaffung. Beim Kauf einer KI-Lösung wird eine bereits entwickelte Software von einem Drittanbieter lizenziert oder als Service genutzt. Diese Option ermöglicht eine schnelle Implementierung und reduziert interne Entwicklungsaufwände, kann aber Einschränkungen bei Anpassungen und Abhängigkeiten von externen Anbietern mit sich bringen. Zudem müssen Unternehmen sicherstellen, dass die Lösung regulatorische und sicherheitstechnische Anforderungen erfüllt. Buy ist eine gute Wahl, wenn eine bewährte Standardlösung für den Anwendungsfall existiert.

Vorteile:

- Schnellere Implementierung
- Hersteller übernimmt Wartung, Updates und teilweise Compliance-Verantwortung
- Zugang zu bewährten Lösungen mit etablierten Sicherheitsstandards
- Reduzierter Entwicklungsaufwand

Nachteile:

- Abhängigkeit von einem externen Anbieter
- Begrenzte Anpassungsmöglichkeiten
- Langfristige Lizenz- und Wartungskosten
- Risiko von Lieferantenwechseln oder Geschäftsaufgaben des Anbieters

Rechtliche Verantwortlichkeiten:

- Prüfung der Zertifizierung und Konformität des Anbieters gemäß AI-Act
- Vertragliche Absicherung von Transparenz- und Auditrechten
- Einhaltung von KRITIS- und Cybersicherheitsrichtlinien
- Datenschutz- und Sicherheitsgarantien in Service-Level-Agreements (SLAs)

Beispiel:

Ein Krankenhaus entscheidet sich für den Kauf einer zertifizierten KI-Software zur Patientenstrom-Optimierung. Das System erfüllt bereits alle regulatorischen Anforderungen, sodass sich das Krankenhaus auf den Betrieb konzentrieren kann, anstatt eigene KI-Modelle zu entwickeln.

Option 3: Transfer

Die Transfer-Option beinhaltet die Anpassung bestehender Open-Source-KI-Systeme oder die Übernahme und Weiterentwicklung externer Lösungen. Dies kombiniert die Vorteile von Make und Buy, indem es Kosten reduziert und gleichzeitig Anpassungsmöglichkeiten bietet. Die Herausforderung besteht in der technischen Integration und der langfristigen Wartung, insbesondere wenn die Open-Source-Community oder der ursprüngliche Anbieter die Unterstützung einstellt. Diese Option eignet sich für Organisationen, die eine flexible, aber kosteneffiziente Lösung benötigen.

Vorteile:

- Nutzung bereits bewährter Technologien
- Geringerer Entwicklungsaufwand als bei Eigenentwicklung
- Potenzielle staatliche Förderungen oder regulatorische Unterstützung
- Möglichkeit zur Kooperation mit anderen KRITIS-Betreibern

Nachteile:

- Abhängigkeit von regulatorischen und politischen Rahmenbedingungen
- Begrenzte Anpassungsmöglichkeiten
- Interoperabilitätsprobleme mit bestehenden Systemen

Rechtliche Verantwortlichkeiten:

- Sicherstellung der Kompatibilität mit nationalen Sicherheitsvorgaben
- Berücksichtigung bilateraler oder europäischer Vertragsbedingungen
- Kontrolle über Datenhoheit und Zugriff durch Dritte

Beispiel:

Eine europäische Verkehrsinfrastrukturbehörde übernimmt eine KI-basierte Sicherheitslösung aus einem anderen EU-Land im Rahmen eines Technologie-Transfer-Abkommens. Dies spart Entwicklungskosten, erfordert jedoch Anpassungen an nationale Anforderungen.

Fazit

Die Wahl zwischen Make, Buy oder Transfer hängt von strategischen, technischen und regulatorischen Faktoren ab. Organisationen sollten regelmäßige Risikobewertungen durchführen, um die sicherste und effizienteste Option zu wählen.

Empfohlene Schritte:

1. Erstellung einer detaillierten Kosten-Nutzen-Analyse für jede Option
2. Einbindung von Fachexperten aus IT, Recht und Cybersicherheit
3. Berücksichtigung von langfristigen Wartungs- und Compliance-Anforderungen
4. Verhandlung vertraglicher Absicherungen mit Anbietern oder Partnern

Nur durch eine sorgfältige Planung und eine vorausschauende Strategie kann sichergestellt werden, dass kritische Infrastrukturen zukunftssicher, resilient und gesetzeskonform bleiben.

Handlungsempfehlungen

Die zunehmende Integration von Künstlicher Intelligenz (KI) in kritische Infrastrukturen (KRITIS) bietet enorme Chancen, bringt aber auch erhebliche Herausforderungen mit sich. Während KI dazu beitragen kann, Prozesse zu optimieren, Sicherheit zu erhöhen und betriebliche Effizienz zu steigern, sind gleichzeitig Risiken wie Sicherheitslücken, Datenschutzprobleme und regulatorische Hürden zu berücksichtigen.

Der EU AI-Act und verwandte Regularien setzen klare Rahmenbedingungen für den sicheren Einsatz von KI in Hochrisiko-Bereichen, zu denen viele KRITIS-Sektoren gehören. Unternehmen und Organisationen müssen daher nicht nur technische, sondern auch rechtliche und ethische Anforderungen erfüllen. Eine durchdachte Make-Buy-or-Transfer-Strategie kann helfen, fundierte Entscheidungen über die Entwicklung, Beschaffung oder Anpassung von KI-Systemen zu treffen.

Die zentrale Erkenntnis ist, dass KI kein Selbstzweck ist. Die Implementierung muss sich an konkreten Anwendungsfällen orientieren, bei denen KI einen messbaren Mehrwert bietet, ohne dabei grundlegende Sicherheits- oder Compliance-Risiken einzugehen. Der Einsatz muss zudem kontinuierlich überwacht und angepasst werden, um technologische Entwicklungen, regulatorische Änderungen und neue Bedrohungslagen zu berücksichtigen.

Handlungsempfehlungen

1. **Strategische Planung und Bedarfsanalyse**

- Organisationen sollten vor der Einführung von KI eine detaillierte Bedarfsanalyse durchführen, um zu ermitteln, wo KI tatsächlich einen Mehrwert bietet.
- Dabei sollte eine klare Kosten-Nutzen-Abwägung zwischen Eigenentwicklung, Kauf oder Anpassung bestehender Systeme erfolgen.

2. **Einhaltung regulatorischer Anforderungen**

- KI-Systeme in KRITIS müssen strikt den Vorgaben des EU AI-Acts sowie relevanter Datenschutz- und Sicherheitsvorschriften entsprechen.
- Unternehmen sollten frühzeitig regulatorische Anforderungen prüfen und gegebenenfalls externe Experten hinzuziehen, um Compliance sicherzustellen.

3. **Sicherheits- und Risikomanagement priorisieren**

- KI-Systeme müssen auf Cybersicherheitsrisiken überprüft werden, insbesondere im Hinblick auf Manipulation, adversarielle Angriffe und unerwartetes Verhalten.
- Kritische Infrastrukturen sollten regelmäßige Sicherheitsüberprüfungen und Audits für KI-Modelle einführen.

4. **Transparenz und Nachvollziehbarkeit sicherstellen**

- Hochrisiko-KI-Systeme müssen nachvollziehbare Entscheidungsprozesse haben, um Vertrauen und Compliance sicherzustellen.
- Unternehmen sollten Dokumentationen und Erklärbarkeitstools nutzen, um sicherzustellen, dass KI-Entscheidungen überprüfbar sind.

5. **Schulung und Bewusstseinsbildung**

- Mitarbeiter müssen auf den Umgang mit KI geschult werden, insbesondere in Bezug auf potenzielle Risiken und regulatorische Anforderungen.
- Sicherheitsverantwortliche sollten regelmäßig Weiterbildungen zu aktuellen Entwicklungen im Bereich KI-Sicherheit und -Regulierung erhalten.

6. **Langfristige Flexibilität sicherstellen**

- Da sich die technologische Landschaft schnell verändert, sollten KI-Strategien langfristig flexibel gestaltet werden.
- Verträge mit Anbietern sollten Klauseln enthalten, die Updates, Sicherheitspatches und Weiterentwicklungen regeln.

7. **Make-Buy-or-Transfer-Entscheidungen strukturiert treffen**

- Ein standardisierter Beschaffungsleitfaden für KI hilft dabei, systematische Entscheidungen über Eigenentwicklung, Kauf oder Anpassung externer Lösungen zu treffen.
- Checklisten und Risikoanalysen sollten vor jeder Implementierung eingesetzt werden, um die beste Strategie zu identifizieren.

Fazit: KI mit Verantwortung gestalten

KI kann eine transformative Rolle in kritischen Infrastrukturen spielen, aber nur, wenn sie mit Bedacht, Sicherheit und strategischer Weitsicht implementiert wird. Unternehmen müssen nicht nur technologische, sondern auch rechtliche und ethische Herausforderungen meistern, um die Vorteile von KI verantwortungsvoll zu nutzen. Handlungsempfehlungen wie eine klare Strategie, regulatorische Compliance und Sicherheitsmaßnahmen sind entscheidend, um das Potenzial der KI in KRITIS zu maximieren und gleichzeitig Risiken zu minimieren.

6 Make-Buy-or Transfer Guide für KI

Die folgende vereinfachte Entscheidungsstruktur hilft, die bestmögliche Lösung für eine bestimmte Organisation oder Infrastruktur zu identifizieren:

1. **Technische Kapazitäten:** Verfügt die Organisation über eigene F&E-Ressourcen und IT-Fachkräfte?
2. **Regulatorische Anforderungen:** Gibt es gesetzliche Vorgaben zu Datenschutz, Datensouveränität oder Sicherheitsstandards?
3. **Anpassungsbedarf:** Benötigt die Lösung spezifische Anpassungen oder individuelle Konfigurationen?
4. **Marktverfügbarkeit:** Gibt es etablierte Lösungen am Markt, die die Anforderungen erfüllen?
5. **Kosten-Nutzen-Analyse:** Ist die langfristige Wartung und Weiterentwicklung wirtschaftlich tragfähig?

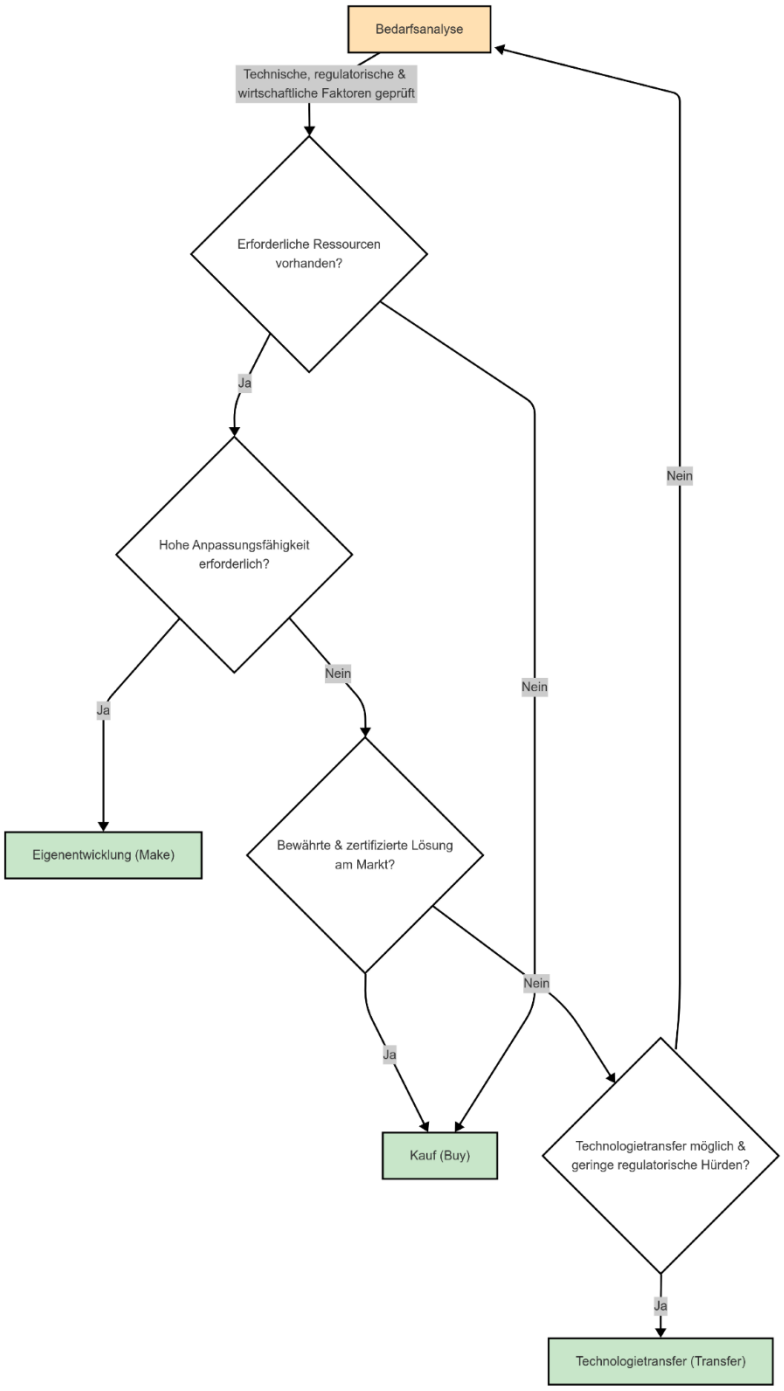


Abbildung 42: Ablauf

Falls die Organisation starke interne Kapazitäten hat und hohe Anpassbarkeit erfordert, kann eine Eigenentwicklung sinnvoll sein. Ist hingegen eine schnelle Umsetzung erforderlich und eine zuverlässige Lösung am Markt vorhanden, ist der Kauf vorzuziehen. Falls eine Technologie bereits existiert und durch Kooperation oder staatliche Unterstützung erworben werden kann, könnte eine Transferlösung die beste Wahl sein.

Der Entscheidungsbaum ist folgendermaßen strukturiert:

- Falls eine Organisation über hohe IT-Kompetenz, regulatorische Kontrolle und spezifische Anpassungsbedarfe verfügt, ist **Make** die bevorzugte Option.
- Wenn eine bewährte Lösung am Markt verfügbar ist, die alle Anforderungen erfüllt, sollte **Buy** in Betracht gezogen werden.
- Falls eine bereits existierende, erprobte Technologie von einem Partner übernommen werden kann und regulatorische Hürden gering sind, bietet sich **Transfer** als geeignete Lösung an.

Option 1: Make

Die Option 1: Make oder Eigenentwicklung einer KI-Lösung bedeutet, dass eine Organisation die Software intern entwickelt, anpasst und betreibt. Dies bietet maximale Kontrolle über Sicherheit, Datenschutz und Anpassungsfähigkeit an spezifische Anforderungen. Allerdings erfordert diese Option hohe Investitionen in Entwicklung, Fachkräfte und Wartung. Sie ist besonders sinnvoll, wenn bestehende Lösungen nicht die erforderlichen Sicherheits- oder Compliance-Standards erfüllen.

Vorteile:

- Volle Kontrolle über Technologie, Daten und Anpassungen
- Keine Abhängigkeit von externen Anbietern
- Möglichkeit, sicherheitskritische Anforderungen maßgeschneidert zu erfüllen
- Langfristig potenziell kosteneffizienter bei nachhaltiger Nutzung

Nachteile:

- Hoher initialer Entwicklungsaufwand
- Erfordert spezialisierte IT- und Sicherheitsexperten
- Wartung und Weiterentwicklung sind langfristige Verpflichtungen
- Regulatorische Verantwortung für Datenschutz, Sicherheit und Transparenz

Rechtliche Verantwortlichkeiten:

- Einhaltung des AI-Act, insbesondere Art. 9 (Risikomanagement) und Art. 27 (Grundrechte-Folgenabschätzung)
- Sicherstellung der DSGVO-Konformität
- Dokumentations- und Auditpflichten für behördliche Kontrollen
- Haftung für Fehlentscheidungen des KI-Systems

Beispiel:

Ein nationaler Netzbetreiber entscheidet sich, ein eigenes KI-basiertes Frühwarnsystem für Cyberangriffe zu entwickeln. Dies ermöglicht maximale Kontrolle über Algorithmen und Datenverarbeitung, erfordert jedoch eine erhebliche Investition in Forschung und Entwicklung sowie laufende Compliance-Prüfungen.

Option 2: Buy

Die Option 2: Buy ist Fremdbeschaffung. Beim Kauf einer KI-Lösung wird eine bereits entwickelte Software von einem Drittanbieter lizenziert oder als Service genutzt. Diese Option ermöglicht eine schnelle Implementierung und reduziert interne Entwicklungsaufwände, kann aber Einschränkungen bei Anpassungen und Abhängigkeiten von externen Anbietern mit sich bringen. Zudem müssen Unternehmen sicherstellen, dass die Lösung regulatorische und sicherheitstechnische Anforderungen erfüllt. Buy ist eine gute Wahl, wenn eine bewährte Standardlösung für den Anwendungsfall existiert.

Vorteile:

- Schnellere Implementierung
- Hersteller übernimmt Wartung, Updates und teilweise Compliance-Verantwortung
- Zugang zu bewährten Lösungen mit etablierten Sicherheitsstandards
- Reduzierter Entwicklungsaufwand

Nachteile:

- Abhängigkeit von einem externen Anbieter
- Begrenzte Anpassungsmöglichkeiten
- Langfristige Lizenz- und Wartungskosten
- Risiko von Lieferantenwechseln oder Geschäftsaufgaben des Anbieters

Rechtliche Verantwortlichkeiten:

- Prüfung der Zertifizierung und Konformität des Anbieters gemäß AI-Act
- Vertragliche Absicherung von Transparenz- und Auditrechten
- Einhaltung von KRITIS- und Cybersicherheitsrichtlinien
- Datenschutz- und Sicherheitsgarantien in Service-Level-Agreements (SLAs)

Beispiel:

Ein Krankenhaus entscheidet sich für den Kauf einer zertifizierten KI-Software zur Patientenstrom-Optimierung. Das System erfüllt bereits alle regulatorischen Anforderungen, sodass sich das Krankenhaus auf den Betrieb konzentrieren kann, anstatt eigene KI-Modelle zu entwickeln.

Option 3: Transfer

Die Transfer-Option beinhaltet die Anpassung bestehender Open-Source-KI-Systeme oder die Übernahme und Weiterentwicklung externer Lösungen. Dies kombiniert die Vorteile von Make

und Buy, indem es Kosten reduziert und gleichzeitig Anpassungsmöglichkeiten bietet. Die Herausforderung besteht in der technischen Integration und der langfristigen Wartung, insbesondere wenn die Open-Source-Community oder der ursprüngliche Anbieter die Unterstützung einstellt. Diese Option eignet sich für Organisationen, die eine flexible, aber kosteneffiziente Lösung benötigen.

Vorteile:

- Nutzung bereits bewährter Technologien
- Geringerer Entwicklungsaufwand als bei Eigenentwicklung
- Potenzielle staatliche Förderungen oder regulatorische Unterstützung
- Möglichkeit zur Kooperation mit anderen KRITIS-Betreibern

Nachteile:

- Abhängigkeit von regulatorischen und politischen Rahmenbedingungen
- Begrenzte Anpassungsmöglichkeiten
- Interoperabilitätsprobleme mit bestehenden Systemen

Rechtliche Verantwortlichkeiten:

- Sicherstellung der Kompatibilität mit nationalen Sicherheitsvorgaben
- Berücksichtigung bilateraler oder europäischer Vertragsbedingungen
- Kontrolle über Datenhoheit und Zugriff durch Dritte

Beispiel:

Eine europäische Verkehrsinfrastrukturbehörde übernimmt eine KI-basierte Sicherheitslösung aus einem anderen EU-Land im Rahmen eines Technologie-Transfer-Abkommens. Dies spart Entwicklungskosten, erfordert jedoch Anpassungen an nationale Anforderungen.

Fazit

Die Wahl zwischen Make, Buy oder Transfer hängt von strategischen, technischen und regulatorischen Faktoren ab. Organisationen sollten regelmäßige Risikobewertungen durchführen, um die sicherste und effizienteste Option zu wählen.

Prozessübersicht

Die Checkliste und das Scoring-Modell sind zwei eng miteinander verbundene Werkzeuge, die zusammen einen strukturierten und transparenten Entscheidungsprozess ermöglichen.

1. **Die Checkliste als Vorbereitung (siehe Kapitel 8)**

- Die Checkliste dient dazu, alle relevanten Aspekte eines KI-Projekts systematisch zu erfassen. Sie führt den Anwender Schritt für Schritt durch Bereiche wie Anforderungsanalyse, Stakeholder-Einbindung, Risikomanagement, Ausschreibung und Vertragsgestaltung.
- Dabei werden sowohl technische, organisatorische als auch rechtliche Kriterien erfasst. Diese Checkliste stellt sicher, dass nichts übersehen wird und bildet die Basis für die spätere Bewertung.
- Checkliste = Vollständige Datensammlung und Vorbereitung.

2. **Das Scoring-Modell für die strategische Entscheidungsfindung (siehe Kapitel 9)**

- Das Scoring-Modell baut auf den in der Checkliste erfassten Kriterien auf und quantifiziert diese in Form von Punktzahlen (auf einer Skala von 1 bis 10) – jeweils für die Optionen „Make“, „Buy“ oder „Transfer“.
- Jedes Kriterium wird zudem gewichtet, um seine relative Bedeutung im Gesamtprozess darzustellen. Die gewichteten Punktzahlen werden dann addiert, um herauszufinden, welche Option im konkreten Fall am besten passt.
- Scoring-Modell = Quantitative Bewertung und Vergleich der Alternativen

3. **So hängen beide zusammen und wie sie verwendet werden:**

- **Erfassung und Vorbereitung:** Zuerst füllt der Anwender die Checkliste aus, um alle relevanten Informationen und Anforderungen zu sammeln.
- **Bewertung:** Anschließend nutzt er diese Informationen, um im Scoring-Modell die jeweiligen Optionen zu bewerten. Dabei werden die in der Checkliste identifizierten Aspekte (z. B. Budget, technische Expertise, Zeitrahmen etc.) in Punkte umgewandelt.
- **Transparente Entscheidungsfindung:** Durch die Kombination aus der umfassenden Checkliste und dem objektiven Scoring-Modell erhält man eine klare, nachvollziehbare Entscheidungsgrundlage, die zeigt, ob es sinnvoller ist, eine Lösung selbst zu entwickeln, extern zu beschaffen oder bestehende Komponenten zu adaptieren.

Checkliste

Mit den Checklisten und Scoring-Modellen erhält die Organisation ein umfassendes Instrumentarium, den gesamten Beschaffungsprozess detailliert zu planen, zu bewerten und kontinuierlich zu optimieren. Die strukturierte Vorgehensweise, die Einbindung aller relevanten Kriterien und die transparente Dokumentation bilden die Grundlage für fundierte Entscheidungen und

tragen maßgeblich dazu bei, dass die beschaffte KI-Lösung den hohen Anforderungen kritischer Infrastrukturen entspricht und zukunftssicher implementiert werden kann.

Erweiterte Checkliste für die Anforderungsanalyse

Detaillierte Erfassung der Geschäftsziele und Use Cases

- **Geschäftsziele und strategische Vision:**
  - Formulierung der übergeordneten Ziele, z. B. Effizienzsteigerung, Erhöhung der Sicherheit oder Verbesserung der Reaktionszeiten in kritischen Infrastrukturen.
  - Festlegung, wie die KI-Lösung zur Erreichung dieser Ziele beitragen soll.
  - Identifizierung von Use Cases, die direkt mit den strategischen Zielen verknüpft sind (z. B. Überwachung von Sicherheitssystemen, automatisierte Alarmierung bei Störungen).
- **Spezifische Anwendungsfälle:**
  - Detaillierte Beschreibung der geplanten Use Cases, inkl. Szenarien, in denen die KI-Lösung operieren soll.
  - Festlegung von Erfolgskriterien und messbaren KPIs (Key Performance Indicators).
  - Analyse der benötigten Funktionalitäten und Schnittstellen zu bestehenden Systemen.

Technische und organisatorische Anforderungen

- **Technische Anforderungen:**
  - Performance und Skalierbarkeit: Welche Datenmengen müssen verarbeitet werden? Wie hoch sind die Echtzeitanforderungen?
  - Sicherheitsanforderungen: Welche Verschlüsselungsstandards, Authentifizierungsverfahren und Zugriffskontrollen sind erforderlich?
  - Integrationsfähigkeit: Welche APIs oder Schnittstellen müssen vorhanden sein, um die Lösung in die bestehende IT-Landschaft einzubinden?
  - Zukunftssicherheit: Wie flexibel ist die Lösung hinsichtlich zukünftiger Erweiterungen und neuer Technologien (z. B. Edge Computing, IoT-Anbindungen)?
- **Organisatorische Anforderungen:**
  - Zuständigkeiten und Rollen: Welche Abteilungen und welche Ansprechpartner sind in den Beschaffungsprozess eingebunden?
  - Schulungsbedarf: Welche internen Schulungen und Trainings sind erforderlich, um die KI-Lösung effektiv zu nutzen?
  - Change-Management: Wie wird der Übergang von der alten zur neuen Lösung geplant und begleitet?
  - Kommunikations- und Eskalationsprozesse: Welche internen Kommunikationswege und Eskalationsprozesse sind zu etablieren, um bei Problemen schnell reagieren zu können?

Rechtliche und Compliance-Vorgaben

- **Datenschutz und DSGVO:**
  - Konkrete Maßnahmen zur Datenminimierung und Anonymisierung.
  - Regelungen zum Zugriff auf sensible Daten und zur Datenaufbewahrung.
  - Dokumentation der Datenflüsse und der eingesetzten Sicherheitsmaßnahmen.
- **Regulatorische Anforderungen:**
  - Einhaltung von Vorgaben des AI-Act, Data Act und weiteren relevanten Normen.
  - Anforderungen an Zertifizierungen (z. B. ISO/IEC 27001) und externe Audits.
  - Vorgaben zur Haftung und zur vertraglichen Absicherung im Falle von Sicherheitsvorfällen.
- **Interne Richtlinien:**
  - Abgleich der externen Anforderungen mit den internen Richtlinien und Strategien.
  - Festlegung von internen Qualitätsstandards, die über die gesetzlichen Mindestanforderungen hinausgehen.

Stakeholder-Einbindung und Dokumentation

- **Interne Stakeholder:**
  - Identifikation und Einbindung relevanter Abteilungen (IT, Recht, Betrieb, Controlling).
  - Regelmäßige Meetings und Workshops zur Abstimmung der Anforderungen.
  - Dokumentation von Entscheidungen und Änderungswünschen in einem zentralen System.
- **Externe Stakeholder:**
  - Einbindung von externen Beratern und Experten (z. B. aus der Sicherheitsbranche, Rechtsanwälte).
  - Abstimmung mit Regulierungsbehörden und Normungsgremien, um künftige Entwicklungen frühzeitig zu berücksichtigen.
- **Dokumentation:**
  - Erstellung eines umfassenden Lastenhefts, das alle Anforderungen detailliert beschreibt.
  - Führen eines Änderungsprotokolls, das alle Anpassungen und Entscheidungen nachvollziehbar festhält.

Die Checkliste soll sicherstellen, dass bei der Erstellung und Durchführung der Ausschreibung alle wesentlichen Aspekte berücksichtigt werden.

Kriterien für die Entscheidung – Scoring-Modell

Das Scoring-Modell ist ein strukturiertes, quantitatives Bewertungsinstrument, das Unternehmen dabei unterstützt, komplexe Entscheidungen nachvollziehbar und objektiv zu treffen. Insbesondere in Bereichen, in denen mehrere Alternativen – wie zum Beispiel die interne Entwicklung, der externe Kauf oder der Transfer bestehender Lösungen – miteinander verglichen werden müssen, stellt das Scoring-Modell eine wesentliche Grundlage dar, um den Entscheidungsprozess zu systematisieren und zu rationalisieren.

Das Hauptziel des Scoring-Modells ist es, unterschiedliche Handlungsoptionen anhand vordefinierter Kriterien zu bewerten und deren relativen Nutzen zu quantifizieren. Unternehmen stehen oft vor der Herausforderung, dass Entscheidungen nicht nur auf subjektiven Eindrücken beruhen, sondern auch eine Vielzahl von Einflussfaktoren berücksichtigen müssen. Mit einem Scoring-Modell können beispielsweise Kriterien wie Budget, technisches Know-how, Zeithorizont, langfristige Betriebskosten, Compliance-Anforderungen sowie Flexibilität und Anpassungsfähigkeit systematisch erfasst und bewertet werden. Dies führt zu einer transparenten Entscheidungsgrundlage, die es ermöglicht, die optimale Handlungsoption herauszustellen und diese Entscheidung gegenüber internen sowie externen Stakeholdern zu begründen.

Kriterium	Gewichtung	Punktzahl Make	Punktzahl Buy	Punktzahl Transfer
Budget und Ressourcen	25 %			
Technische Expertise	20 %			
Zeithorizont	15 %			
Langfristige Wirtschaftlichkeit	20 %			
Flexibilität und Anpassungsfähigkeit	10 %			
Rechtliche und Compliance-Aspekte	10 %			

Tabelle 43: Hauptkategorien in dem Scoring-Modell

Hinweis: Die tatsächlichen Punktzahlen werden in einem internen Workshop ermittelt und fließen in die Entscheidungsfindung ein. In dem Appendix sieht man die erweiterte Version.

Logischer Aufbau des Scoring-Modells

Der Aufbau eines Scoring-Modells erfolgt in mehreren strukturierten Schritten:

1. **Kriterienfestlegung:**
   Zunächst werden alle relevanten Kriterien identifiziert, die für die Entscheidungsfindung von Bedeutung sind. Diese Kriterien sollten sowohl quantitative als auch qualitative Aspekte abdecken. Quantitative Kriterien können beispielsweise Kosten, Zeitaufwand oder Ressourceneinsatz sein, während qualitative Kriterien Aspekte wie Innovationsgrad, Risiko oder strategische Passung berücksichtigen.

2. **Gewichtung der Kriterien:**
   Nicht alle Kriterien haben den gleichen Stellenwert. Daher werden den einzelnen Kriterien Gewichtungen zugeordnet, die deren relative Bedeutung im Entscheidungsprozess widerspiegeln. Die Summe aller Gewichtungen sollte 100 % ergeben. Eine sorgfältige Gewichtung ist entscheidend, um sicherzustellen, dass kritische Faktoren, wie etwa Compliance oder Sicherheit, angemessen berücksichtigt werden.

3. **Bewertungsskala:**
   Für jedes Kriterium wird eine Bewertungsskala definiert, typischerweise von 1 bis 10, wobei 1 für eine sehr schlechte und 10 für eine sehr gute Erfüllung des Kriteriums steht. Diese Skala ermöglicht eine differenzierte Bewertung der Alternativen und schafft eine Basis für den Vergleich.

4. **Berechnung der gewichteten Punkte:**
   Die einzelnen Bewertungen werden mit der jeweiligen Gewichtung multipliziert. Auf diese Weise entstehen gewichtete Punktzahlen, die den Beitrag jedes Kriteriums zum Gesamtergebnis darstellen. Die Summe dieser gewichteten Punkte für jede Option gibt dann Aufschluss darüber, welche Alternative insgesamt am vorteilhaftesten ist.

5. **Dokumentation und Transparenz:**
   Der gesamte Bewertungsprozess – von der Auswahl der Kriterien über die Gewichtung bis hin zur finalen Punktzahl – sollte detailliert dokumentiert werden. Diese Dokumentation ist wichtig, um die Entscheidungsfindung nachvollziehbar zu machen und im Falle von Audits oder internen Reviews aufzeigen zu können, wie die Entscheidung zustande gekommen ist.

Zusammenfassung und Nutzen der erweiterten Checklisten und Scoring-Modelle

Die ausführlich erweiterten Checklisten und Scoring-Modelle stellen sicher, dass alle wesentlichen Aspekte des Beschaffungsprozesses – von der initialen Anforderungsanalyse über die Ausschreibung bis hin zur Vertragsgestaltung – systematisch erfasst und bewertet werden. Durch diese detaillierte Herangehensweise profitieren Organisationen in mehrfacher Hinsicht:

- **Transparenz:** Jede Entscheidung wird nachvollziehbar dokumentiert. Dies erleichtert interne Reviews und externe Audits gleichermaßen.
- **Objektivität:** Die Bewertung anhand klar definierter Kriterien und Gewichtungen reduziert subjektive Einflüsse und ermöglicht fundierte Entscheidungen.
- **Flexibilität:** Regelmäßige Feedback-Schleifen und die Möglichkeit, die Modelle anzupassen, sorgen dafür, dass der Beschaffungsprozess stets an aktuelle technologische und regulatorische Entwicklungen angepasst werden kann.
- **Risikominimierung:** Durch die detaillierte Erfassung von Sicherheits-, Compliance- und Integrationsanforderungen wird das Risiko von Fehlentscheidungen und unzureichender Umsetzung deutlich reduziert.
- **Effizienzsteigerung:** Klare Prozesse und detaillierte Checklisten ermöglichen eine strukturierte und zeiteffiziente Durchführung des gesamten Beschaffungsprozesses.

Diese erweiterten Instrumente sind somit ein zentraler Baustein, um KI-Lösungen für kritische Infrastrukturen und HighRisk-Anwendungen langfristig erfolgreich und sicher zu beschaffen und zu implementieren.

Zusammenfassung

Das Hauptziel des Projekts bestand in der Analyse der neu entstehenden EU-weiten Regularien in Bezug auf die Nutzung von AI, speziell im Hochrisikobereich, bzw. im Anwendungsgebiet kritischer Infrastrukturen. Daraus wurden Handlungsleitfäden für das BM.I und Betreiber kritischer Infrastrukturen entwickelt. Zusätzlich zu diesen Analysen, wurden im Rahmen einer explorativen Szenarienanalyse analysiert, wie sich das Umfeld künftig entwickeln kann und welche Auswirkungen dies auf die Arbeit der LEAs besitzen wird. Dazu wurde eine prominente Methodik für die Bedürfnisse der ESA in einem sehr rasch wachsenden und sich disruptiv verändernden Forschungsfeld umgebaut. Zusätzlich wurde die Methodik dergestalt verändert, dass sie keinerlei proprietäre Software mehr für ihren Einsatz benötigt, sowie rasch genug durchgeführt werden kann, dass sie auch öfter realistisch eingesetzt werden kann.

Basierend auf den Analysen wurden nicht nur die Use-Cases des Bedarfsträgers bearbeitet, sondern auch neue Werkzeuge entwickelt, die sowohl dem Bedarfsträger als auch Firmen zur Verfügung gestellt werden können: Data-Science-Trajectories erlauben die strukturierte Entwicklung von neuen AI-Anwendung, bzw. deren Wartung und gezielte Änderung. Der Make-or-Buy-Guide hilft Beschaffenden in der Entscheidung, im Lichte der zusätzlichen Rechte und Pflichten, die durch den AI-Act erwachsen, eine Kauf- oder Entwicklungsentscheidung zu treffen. Last, but not least, war auch das Thema des Risikomanagements in AI-Systemen ein wichtiger Aspekt der Analysen dieses Projekts. Durch die Durchführung von Expert*innen-Interviews konnte sichergestellt werden, dass die Erkenntnisse nicht auf rein akademischen Arbeiten basieren, die zum Zeitpunkt der Erfassung in einem sich sehr schnell entwickelnden technologischen Umfeld schon veraltet sein könnten, sondern wichtige Expert*innen auf ihren Gebieten direkt und konkret nach wichtigen Einflussfaktoren, aber auch von ihnen antizipierten neuen Gefahren, sowie Anwendungen, speziell im Umfeld von LEAs, befragt werden konnten.

Zusammengefasst konnten alle Ziele des gegenständlichen Projekts im Rahmen der folgenden wesentlichen Projektergebnisse erreicht werden:

(1) Risikomanagement für AI mit Risikoregister, Fallbeispiel

(2) Detaillierte Rechtlichen Analyse

(3) Explorative Szenarioanalyse

(4) Experteninterview

(5) Beschaffungsleitfadens für sichere AI, Check-Listen

Anhang A – Leere Checkliste

Leere Checkliste (Vorlage)

A. Grundinformationen zum Projekt

Checkpunkt	Beschreibung	Erledigt (Ja/Nein)
Projektname / Use Case	Wie heißt das Projekt bzw. der Anwendungsfall?	
Ziele und Mehrwert	Welche Ziele sollen durch den Einsatz von KI erreicht werden?	
Beteiligte Abteilungen / Teams	Wer ist zuständig (IT, Recht, Security, Fachbereiche)?	
Verantwortliche Personen	Wer ist für die Gesamtprojektleitung zuständig?	

B. Anforderungsanalyse

Checkpunkt	Beschreibung	Erledigt (Ja/Nein)
Fachliche Anforderungen	Welche konkreten Funktionen werden benötigt (z. B. Datenauswertung, Alarme)?	
Technische Anforderungen	Performance, Skalierbarkeit, Schnittstellen etc.	
Organisatorische Anforderungen	Schulungen, Zuständigkeiten, Abläufe	
Rechtliche / Compliance-Anforderungen	DSGVO, AI-Act, Data Act, weitere Vorschriften	
Zielgruppen und Endnutzer	Wer nutzt die KI-Lösung, in welcher Form?	

C. Stakeholder- und Ressourcenanalyse

Checkpunkt	Beschreibung	Erledigt (Ja/Nein)
Interne Stakeholder	Welche Abteilungen / Personen sind beteiligt?	
Externe Stakeholder	Externe Partner, Berater, Lieferanten	
Ressourcenplanung	Budget, personelle Ressourcen, Zeitplanung	
Verfügbarkeit von Fachexpertise	Benötigen wir zusätzliche Schulungen oder externe Expertise?	

D. Risikomanagement und Sicherheitsaspekte

Checkpunkt	Beschreibung	Erledigt (Ja/Nein)
Identifikation von Risiken	Mögliche Risiken (z. B. Datenlecks, Fehlalarme, Diskriminierung)	
Bewertung der Risiken	Eintrittswahrscheinlichkeit, Schadensausmaß	
Gegenmaßnahmen / Notfallpläne	Technische und organisatorische Maßnahmen zur Risikominimierung	
Monitoring und Incident-Response	Kontinuierliche Überwachung, Eskalationswege	

E. Ausschreibung und Anbieterbewertung

Checkpunkt	Beschreibung	Erledigt (Ja/Nein)
Ausschreibungsdokumente	Leistungsbeschreibung, Bewertungskriterien, Zeitplan	
Make, Buy or Transfer-Entscheidung	Anwendung des Make, Buy or Transfer Guides	
Angebotseinholung und Marktanalyse	Auswahl potenzieller Anbieter, Marktübersicht	
Scoring / Bewertung	Kriterien wie technische Kompetenz, Kosten, Referenzen, Sicherheit	
Anbieterpräsentationen / Proof of Concept	Durchführung von Tests, Pilotprojekten oder Demos	

F. Vertragsgestaltung und rechtliche Absicherung

Checkpunkt	Beschreibung	Erledigt (Ja/Nein)
Vertragsinhalte	Umfang, SLAs, Wartung, Updates, Leistungsbeschreibung	
Datenschutz und DSGVO	Spezifikation von Verantwortlichkeiten, Löschkonzept, Datentransparenz	
Haftung und Gewährleistung	Regelungen für Schadensfälle, Ausfallzeiten, Haftungsbeschränkungen	
Lizenz- und Nutzungsrechte	Festlegung, wem die Rechte an der KI-Lösung gehören	

G. Implementierung und Integration

Checkpunkt	Beschreibung	Erledigt (Ja/Nein)
Projektplan und Meilensteine	Konkrete Zeitplanung, Zuständigkeiten, Phasen	
Integration in bestehende Systeme	Schnittstellen, Kompatibilität, Datenformate	
Schulung der Mitarbeiter	Trainingsprogramme, Dokumentation, Ansprechpartner	
Test- und Pilotphase	Ablauf, Erfolgskriterien, Freigabeprozesse	

H. Betrieb, Monitoring und Weiterentwicklung

Checkpunkt	Beschreibung	Erledigt (Ja/Nein)
Regelmäßige Updates und Wartung	Patch-Management, Sicherheitsupdates	
Performance-Monitoring	Überwachung der KI-Leistung, Alarmierung bei Auffälligkeiten	
Compliance und Audits	Interne / externe Überprüfungen, Dokumentation	
Skalierung und Weiterentwicklung	Anpassung an wachsende Anforderungen, neue Features	

Ausfüllhilfe für das Scoring-Modell

1. **Kriterien definieren und priorisieren:**

- Lege zunächst fest, welche Kriterien für Deine Organisation bzw. Dein Projekt besonders wichtig sind. Typische Kriterien sind Budget, Ressourcen, technisches Know-how, Zeithorizont, Compliance, etc.
- Bestimme für jedes Kriterium eine Gewichtung (in Prozent), wobei die Summe aller Gewichtungen 100 % ergeben sollte.

2. **Punktevergabe (Skala 1–10):**

- Bewerte jede Option (Make, Buy, Transfer) pro Kriterium mit Punkten zwischen 1 (sehr schlecht) und 10 (sehr gut).
- Berücksichtige dabei sowohl quantitative Daten (z. B. konkrete Kosten, Zeitrahmen) als auch qualitative Einschätzungen (z. B. wie gut passt die Lösung zum Geschäftsmodell?).

3. **Gewichtete Punktzahl berechnen:**

- Multipliziere die Punkte jeder Option mit der festgelegten Gewichtung.
- Beispiel: 8 Punkte × 25 % = 2,0.

4. **Summieren und vergleichen:**

- Addiere alle gewichteten Punktzahlen für jede Option.
- Die höchste Gesamtsumme weist auf die für Deine Organisation vorteilhafteste Beschaffungsstrategie hin.

5. **Dokumentation und Transparenz:**

- Halte alle Bewertungen und Begründungen schriftlich fest, um später nachvollziehen zu können, warum Du bestimmte Punkte vergeben hast.
- Diese Dokumentation ist besonders wertvoll, wenn das Projekt in Audits oder Management-Reviews überprüft wird.

6. **Review und Aktualisierung:**

- Das Scoring-Modell kann an Veränderungen angepasst werden (z. B. neue technologische Entwicklungen, geänderte Budgets).
- Führe regelmäßige Reviews durch, insbesondere wenn sich Rahmenbedingungen oder Projektziele ändern.

Bewertungstabellen für

- Grundlegende Kriterien
- Technische Bewertung
- Operative und Rechtliche Bewertung.

Tabelle 3: Bewertung der grundlegenden Kriterien

Kriterium	**Gewichtung (%)**	**Beschreibung**	**Punkte Make (1–10)**	**Punkte Buy (1–10)**	**Punkte Transfer (1–10)**
Budget und Ressourcen	25	Verfügbarkeit interner Ressourcen und Budget zur Eigenentwicklung			

- Investitionskosten	10	Initiale Kosten im Vergleich zu externen Angeboten			
- Betriebskosten und Wartungsaufwand	15	Langfristige Kosten für Betrieb und Support			
Technische Expertise und Know-how	20	Verfügbarkeit von internem Fachwissen für Entwicklung und Integration			
Zeithorizont	15	Dringlichkeit der Lösung und Marktdynamik			
- Implementierungszeit	10	Zeit bis zur vollständigen Inbetriebnahme			
- Flexibilität im Projektverlauf	5	Anpassungsfähigkeit an sich ändernde Anforderungen			
Langfristige Wirtschaftlichkeit	20	Gesamtkostenbetrachtung über den gesamten Lebenszyklus			
- ROI (Return on Investment)	10	Erwarteter wirtschaftlicher Nutzen im Vergleich zu den Kosten			
- Skalierbarkeit und Zukunftssicherheit	10	Möglichkeit, die Lösung an zukünftige Anforderungen anzupassen			
Rechtliche und Compliance-Aspekte	10	Erfüllung aller gesetzlichen Vorgaben und minimierte Haftungsrisiken			
Flexibilität und Anpassungsfähigkeit	10	Fähigkeit der Lösung, sich an neue Anforderungen anzupassen			

Hinweis: Jede Option wird anhand der Punktzahl bewertet (Skala 1–10). Die Gesamtpunktzahl ergibt sich aus der gewichteten Summe der einzelnen Kriterien.

Tabelle 4: Technische Bewertung

Technisches Kriterium	Gewichtung (%)	Beschreibung	Punkte Make	Punkte Buy	Punkte Transfer
Sicherheitsanforderungen	30	Erfüllung von Sicherheitsstandards (z. B. Verschlüsselung, Zugriffskontrollen)			
- Implementierung von Sicherheitsprotokollen	15	Vorhandensein und Qualität von Sicherheitsprotokollen und Maßnahmen			
- Kontinuierliches Monitoring und Auditierung	15	Fähigkeit der Lösung, Sicherheitsvorfälle frühzeitig zu erkennen			
Integrationsfähigkeit	25	Nahtlose Einbindung in bestehende IT-Infrastrukturen			
- Schnittstellen und API-Kompatibilität	15	Verfügbarkeit standardisierter Schnittstellen und APIs			
- Anpassbarkeit an bestehende Systeme	10	Leichtigkeit der Integration und Anpassung an individuelle Anforderungen			
Leistung und Skalierbarkeit	25	Fähigkeit, hohe Datenmengen und Echtzeitanforderungen zu bewältigen			
- Verarbeitungsgeschwindigkeit	15	Geschwindigkeit der Datenverarbeitung und Reaktionszeiten			
- Erweiterbarkeit	10	Möglichkeiten, die Lösung bei steigenden Anforderungen auszubauen			
Innovationsgrad und Zukunftssicherheit	20	Technologischer Stand und zukünftige Weiterentwicklungsmöglichkeiten			
- Anpassungsfähigkeit an neue Technologien	10	Bereitschaft und Fähigkeit, zukünftige Technologien zu integrieren			
- Innovationspotenzial	10	Grad an Innovation und Differenzierung gegenüber Standardlösungen			

Tabelle 5: Operative und rechtliche Bewertung

Operatives/ Rechtliches Kriterium	Gewichtung (%)	Beschreibung	Punkte Make	Punkte Buy	Punkte Transfer
Compliance und rechtliche Anforderungen	40	Erfüllung aller gesetzlichen und regulatorischen Vorgaben			
- Datenschutz und DSGVO-Konformität	20	Maßnahmen zum Schutz personenbezogener Daten und Einhaltung der DSGVO			
- Einhaltung von AI-spezifischen Regelungen	20	Berücksichtigung von Vorgaben aus dem AI-Act, Data Act und weiteren Regularien			
Organisatorische Integration und Support	30	Qualität und Umfang der internen und externen Support-Strukturen			
- Schulungs- und Trainingsprogramme	15	Umfang und Regelmäßigkeit der Mitarbeiterschulungen			
- Change-Management und Supportprozesse	15	Effizienz der internen Abläufe und Unterstützung bei der Implementierung			
Wirtschaftlichkeit und ROI	30	Gesamtkostenbetrachtung und erwarteter Return on Investment			
- Initialkosten vs. langfristige Betriebskosten	15	Vergleich der Investitions- und Betriebskosten			
- Nachhaltiger wirtschaftlicher Nutzen	15	Langfristige Effizienzsteigerungen und Kosteneinsparungen durch die Lösung			

Hinweis: Die Punktzahlen in den Tabellen können durch Workshops mit internen Experten gemeinsam festgelegt und gewichtet werden. Die Gesamtsummen ergeben ein objektives Bild, das als Basis für die finale Entscheidung dient.

Anhang B - Risiko Akteure

Nationale Akteure

Unter dem Begriff „Nationale Akteure“ versteht man staatlich unterstützte und/oder gesponsorte Organisationen, die entsprechend dem Interesse des Herkunftslandes im digitalen Raum operieren, und sich Angriffe auf andere Länder oder Organisationen zum Ziel setzen. Durch direkte oder indirekte staatliche Unterstützung sind solche Gruppen oftmals sehr potent, und verfügen über ein deutlich höheres Niveau an Ressourcen als staatenlose Hackergruppen. Ihre Ziele und Motive decken sich in der Regel mit nationalen Interessen, wobei hier oftmals politische Einflussnahme, wirtschaftliche Vorteile, militärische Überlegenheit, und die nationale Sicherheit genannt wird.

Nationale Akteure sind eine ernstzunehmende Bedrohung, da sie aufgrund ihrer staatlichen Förderung in der Lage sind langanhaltende und komplexe Operationen durchzuführen. Sie sind ein Baustein der strategischen Kriegsführung, und haben in der Regel auch Zugriff auf potente „Cyberwaffen“ und Tools, die in ihrer Kapazität herkömmliche Hacking-Anwendungen weit übersteigen können. Ein erhebliches Problem nationaler Akteure ist darüber hinaus, dass sie in ihrem Herkunftsland rechtliche Immunität genießen können, wodurch ein legales Vorgehen gegen solche Organisationen erschwert wird. Besonders Länder wie Russland, China, und Nordkorea sind dafür bekannt, viele solcher staatlich gesponsorten Hackergruppen zu beherbergen.

Ein Beispiel für solch einen nationalen Akteur ist die „Einheit 61398“ der chinesischen Volksbefreiungsarmee, die eine lange Geschichte an (Industrie)-Spionage und Cyberangriffen verzeichnen kann. Mit dem Ziel der wirtschaftlichen und militärischen Stärkung Chinas durch Diebstahl von Informationen und geistigem Eigentum ist sie eine der aktivsten Gruppen ihrer Art. Besonders die Sektoren Raumfahrt, IT, Verteidigung, aber auch Energie, Finanzdienstleistungen, und Telekommunikation zählen zu den am meisten betroffenen. Von einfachen Phishing-Angriffen, Zero-Day-Exploits wird von dieser Gruppe nahezu jedes Mittel eingesetzt.

Cyberkriminelle

Hinter dem Begriff der Cyberkriminellen steckt das digitale Pendant zu regulären Kriminellen. Der Motivationsgrund für diese Gruppe sind in der Regel finanzielle Gewinne, beispielsweise durch den Verkauf gestohlener Daten, oder persönliche Vorteile. Umfang und Schweregrad der ausgeübten Angriffe sind abhängig von einer Vielzahl an Parametern, wie beispielsweise die Größe der Gruppe, ihre Hacking-Erfahrung, verfügbare Ressourcen, etc.

Bekannte und häufig auftretende Arten von Angriffen sind unter anderem Ransomware, Verletzung der Datensicherheit, DDoS, und Social Engineering. Die Übernahme von existierenden Benutzeraccounts dient dabei oft als erster Schritt, um einen erfolgreichen Angriff zu starten. Dabei spielen schwache Passwörter und Standardkonfigurationen den Cyberkriminellen in die Hände. Ebenso wird legitime Software für beispielsweise Fernzugriffe verwendet, um damit das Einfallstor zu öffnen.

Das Auftreten neuer Technologien und Fortschritte im Bereich der Künstlichen Intelligenz führen zu neuen Bedrohungen. Einerseits können mit Deepfakes und KI-generierten Stimmen die Opfer von Cyberkriminellen bei Social Engineering Angriffen leichter getäuscht werden, andererseits kann KI auch genutzt werden, um Angriffe selbst effizienter zu gestalten und Ressourcen zu sparen. Generierte Nachrichten und Code für Schadsoftware können bereits für eine Vielzahl von Angriffen auch ohne vertiefendes technisches Know-How genutzt werden.

Insider

Als Insider werden Personen bezeichnet, die mit ihren Aktivitäten innerhalb eines Systems oder einer Organisation Schaden anrichten. Durch legitime Zugriffsberechtigungen und Wissen über die Systemlandschaft ist es für sie oft einfacher intern vorzudringen als für externe Angreifer. Insider werden in der Regel in zwei Kategorien aufgeteilt. Einerseits gibt es böswillige Akteure die aktiv gegen die Interessen einer Organisation arbeiten, Daten stehlen, Industriespionage betreiben, oder die Systeme auf andere Arten zu sabotieren. In diesen Fällen ist die Motivationsgrundlage für das aktiv böswillige Handeln ein möglicher finanzieller Vorteil, Rache und Unzufriedenheit, oder aber auch ideologische Überzeugung.

Die zweite Kategorie ist der „unbeabsichtigte" Insider. Hierbei handelt es sich um Personen die aufgrund von einer Mischung aus Fahrlässigkeit, Nachlässigkeit, Unwissenheit, Ignoranz, oder mangelnden Fachkompetenzen durch ihr Handeln Sicherheitsverstöße, Datenverlust, und eine Schwächung der Systeme verursachen können. Häufige Fehlerquellen sind hier das Öffnen von Phishing-Mails, die Verwendung von schwachen Passwörtern und ein unsicherer Umgang mit Login-Informationen, sowie eine falsche Handhabung vertraulicher Daten.

Hacktivisten und Cyberterroristen

Hacktivisten sind Personen oder Gruppen, die ihre technischen Fähigkeiten nutzen, um auf bestimmte Themen aufmerksam zu machen, oder ausgehend von politischer Motivation aktiv werden. Hacktivisten geht es in der Regel darum, auf ein bestimmtes politisches, soziales, oder ideologisches Anliegen hinzuweisen, und sie sehen ihre Angriffe als eine Art des Protestes und Ausdrucks ihrer Standpunkte. Oftmals angegebene Gründe sind beispielsweise die Erhaltung der Freiheit des Internets, Menschenrechte, Umweltschutz, Antikapitalismus, oder aber auch geopolitische Ereignisse wie die russische Invasion der Ukraine.

Da es sich bei Hacktivisten um freiwillige Individuen oder Gruppen handelt, sind ihre finanziellen Mittel und Ressourcen auch dementsprechend überschaubar. Das hat Einfluss auf ihre Möglichkeiten und Reichweite. So sind beispielsweise Denial-of-Service-Angriffe, Website-Defacement, die Veröffentlichung von geheimen Informationen und Daten, sowie das Doxing beliebte Vorgehensweisen.

Eine der wohl bekanntesten Hacktivismus-Gruppen ist Anonymous, eine dezentrale Hacking-Community mit dem generellen Ziel eines „Kampfes gegen korrupte Entitäten", die auch nicht davor zurückschrecken ganze Staaten anzugreifen. Bekannt sind sie unter anderem durch ihre Unterstützung von Protestbewegungen in autoritären Staaten, Angriffe auf diverse PolitikerInnen, und ihre Guy-Fawkes Masken, hinter denen die Anonymität der beteiligten Personen gewahrt werden soll. Darüber hinaus befindet sich das Kollektiv seit der russischen Invasion der Ukraine in einem selbsternannten „Cyber-Krieg" gegen die Russische Föderation, gefolgt von unzähligen Operationen wie der Offenlegung von persönlichen Informationen von 120.000 russischen Soldaten in der Ukraine.

Cyberterroristen haben unter anderem ähnliche Motivationsgründe wie auch Hacktivisten (politische und ideologische Überzeugung), sie haben jedoch andere Ziele und nutzen dafür auch abweichende Methoden. Beim Cyberterrorismus geht es oft um die tatsächliche Schädigung/Zerstörung wichtiger Einrichtungen und kritischer Infrastruktur, darüber hinaus ist auch die Verbreitung von Angst in der Bevölkerung und damit verbundenes Auftreten von Instabilität ein

gewünschtes Ergebnis. Ebenso ist die Manipulation von Wahlen, und Verbreitung von Desinformation ein oft genutztes Mittel um die übergeordneten Ziele zu erreichen.

Weiterführende Informationen

- Sailio, Mirko, Outi-Marja Latvala, and Alexander Szanto. "Cyber threat actors for the factory of the future." *Applied Sciences* 10.12 (2020): 4334.
- Center, Mandiant Intelligence. "APT1: Exposing one of China's cyber espionage units." *Mandian.com* (2013).
- NIST SP 800-82r3, Appendix C
- ENISA Threat Landscape 2023, Kapitel 2
- ENISA Artificial Intelligence Cybersecurity Challenges, Kapitel 4.1
- Janczewski, Lech, and Andrew Colarik, eds. *Cyber warfare and cyber terrorism*. IGI Global, 2007. Section 1

Anhang C – Anwendungsfälle von KI und ihre Risiken

Use Case 1: Analyse sozialer Netzwerke

Soziale Medien und Netzwerke sind neben herkömmlichen Medien ein immer wichtiger werdender Schauplatz für die Verbreitung von Informationen und Desinformation. Auch Inhalte, die auf den ersten Blick harmlos wirken, können einen entscheidenden Beitrag zur Radikalisierung in jegliche Richtung leisten. In Fällen der organisierten Kriminalität oder der Terrorismusbekämpfung können die offengelegten Informationen in sozialen Netzwerken dazu genutzt werden, um mögliche Täter und ihr Umfeld näher auszuforschen. Die Nutzung solcher öffentlich zugänglichen Informationen wird auch als OSINT (Open Source INTelligence) bezeichnet. Die Verwendung von graphbasierten Algorithmen spielt in Kombination mit den erlangten Informationen eine große Rolle.

Dieser Anwendungsfall geht davon aus, dass bereits eine aktive Anordnung zur Durchführung von Ermittlungen erlassen wurde. Einsatzzweck und Erhebungsumfeld der polizeilichen Aktivitäten können dabei vielfältig sein.

Risiken des Anwendungsfalls

Datenschutz und rechtliche Risiken und Ethik

Der Einsatz von KI-gesteuerten Tools zur Analyse, Ausforschung, und potenziellen Überwachung der Bevölkerung durch den Staat kann als erheblicher Eingriff in die Privatsphäre gewertet werden. Zusätzlich könnte dadurch ein Abwandern zu anderen Kommunikationskanälen und eine Änderung am Nutzerverhalten von möglichen Zielen bewirkt werden. Was die reguläre Bevölkerung betrifft, so kann die bloße Existenz einer solchen Technologie großen Unmut hervorrufen, da die vom KI-System genutzten Fähigkeiten auch leicht zur anlasslosen Massenüberwachung eingesetzt und missbraucht werden könnten. Darüber hinaus können Fehler oder Bias im System dazu führen, dass unbeteiligte Personen Ziel der automatisierten Überwachung und Ausforschung werden.

Ein korrektes Absichern der vom KI-System erhobenen und ausgewerteten Daten muss ebenfalls erfolgen, um sie vor unbefugten Zugriffen oder gar Datenlecks und Datendiebstahl zu schützen. Einerseits weil es sich hierbei um Beweismaterial der polizeilichen Ermittlungsarbeit handeln kann, andererseits, weil wie bereits erwähnt möglicherweise unbeteiligte Personen ebenfalls Teil der Erhebung wurden.

Falschpositive und Falschnegative

Je umfangreicher eine Untersuchung des Umfelds einer verdächtigen Person durchgeführt wird, kann es dazu führen, dass unbescholtene Individuen fälschlicherweise als verdächtig/befangen vom KI-System eingestuft werden. Dies kann beispielsweise aufgrund von Korrelationen ohne Relevanz vorkommen. Dadurch kann es zu unnötiger Überwachung oder Verdächtigung von Unschuldigen kommen.

Gleichermaßen sind Falschnegative ein Problem, das bei der Nutzung von KI-Systemen während einer Personen-Umfeldanalyse auftreten kann. Basierend auf einer Vielzahl an Gründen, wie ein nicht ausreichendes Training, ein verzerrtes oder Fehlerhaftes Trainingsdatenset, oder aber auch

nicht erkannte Verbindungen beim Durchlaufen des Algorithmus kann es dazu führen, dass tatsächliche Bedrohungen nicht vom System erkannt werden, oder sie falsch zugeordnet werden.

Diskriminierung und Bias

Der Einsatz von KI-Systemen kann dazu führen, dass bestehende Vorurteile und Diskriminierung weiter verstärkt werden. Dadurch könnten bestimmte Personengruppen überproportional vom System zur Überwachung freigegeben werden, und es kann zu haltlosen Verdächtigungen kommen. Besonders durch den Einsatz von graphbasierten Algorithmen könnten bestimmte soziale und kulturelle Gruppen unverhältnismäßig oft zur Zielscheibe gemacht werden.

Datenqualität

Daten aus sozialen Netzwerken, die zu Ermittlungszwecken verwendet werden, könnten mehrfach problematisch sein. Einerseits könnten die Informationen veraltet sein und somit die Realität der aktuellen Situation verzerren. Andererseits könnten unvollständige Angaben dazu führen, dass diese Unvollständigkeit ebenfalls in die Analyse übernommen wird und somit das Endergebnis beeinflusst. Zusätzlich kann ein Mangel an Kontextinformationen, der für eine menschliche Überprüfung vielleicht gegeben wäre, ebenso zu falschen Ergebnissen führen.

Neben Problemen von unsauberen und unvollständigen Daten besteht auch noch das Problem von aktiv erstellten und verbreiteten Falschinformationen. Auf unterschiedlichste Weise könnten Zielpersonen von Ermittlungen bereits im Vorfeld falsche Informationen und Profile erstellt haben, um durch Manipulation und Täuschung der Behörden die Analysevorgänge zu erschweren, oder auf andere Arten zu behindern.

„A3 – AI Act for Austria“

Rechtliche Aspekte

Dr. iur. Max Cornelson, MA (Arbeitsgruppe Rechtsinformatik / Wiener Zentrum für Rechtsinformatik)

Februar 2025

Inhaltsverzeichnis

1. Definition der „Künstlichen Intelligenz“ nach der KI-VO; Methodologisches.....2
2. Der „risikobasierte Ansatz“; Begriff des Risikos; „Schaden“; Einstufung von KI…6
3. Rollenverteilung und Pflichten nach den Vorschriften der KI-VO über Hochrisiko-Systeme.....14
3.1. technische und organisatorische Maßnahmen.....18
3.2. menschliche Aufsicht.....22
3.3. datenbezogene Verantwortung des Betreibers.....29
3.4. Grundrechte-Folgenabschätzung.....38
3.5. Aufzeichnungspflichten/Protokollierung/Dokumentation.....41
3.6. Zusammenarbeit mit dem Anbieter und der Marktüberwachungsbehörde.....43
4. KI-Einsatz in Ermittlungen nach StPO, SPG, SNG.....48
4.1. Rasterfahndung.....53
4.2. biometrische Echtzeit-Fernidentifizierungssysteme.....56
4.3. Systeme zur nachträglichen biometrischen Fernidentifizierung.....64
4.4. Information betroffener Personen, Erläuterung der Entscheidungsfindung.....68
Exkurs: KI-Einsatz und Werturteile.....73
Exkurs: Datenverordnung.....73
Literatur.....76

1. Definition der „Künstlichen Intelligenz“ nach der KI-VO; Methodologisches

Die Verordnung über Künstliche Intelligenz in ihrer finalen Version[1] definiert in Artikel 3 („Begriffsbestimmungen“) Ziffer 1 den Begriff des „**KI-Systems**“ folgendermaßen:

> „'KI-System' [bezeichnet] ein maschinengestütztes System, das für einen in unterschiedlichem Grade **autonomen Betrieb** ausgelegt ist und das nach seiner Betriebsaufnahme **anpassungsfähig** sein kann und das aus den erhaltenen Eingaben **für explizite oder implizite Ziele ableitet**, wie Ausgaben wie etwa Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erstellt werden, die physische oder virtuelle **Umgebungen beeinflussen** können[.]“

Laut Art 96 Abs 1 lit f KI-VO soll die Europäische Kommission diese Definition durch **Leitlinien** näher konkretisieren. Diese wurden Februar 2025 als Entwurf publiziert und umfassen dreizehn Seiten.[2] Sie sind ausdrücklich nicht rechtsverbindlich[3] und in ihrem Anspruch eher bescheiden: „No automatic determination or exhaustive lists of systems that either fall within or outside the definition of an AI system are possible.“[4]

Im ursprünglichen Entwurf[5] der Verordnung wurde der Begriff der künstlichen Intelligenz noch durch die im Anhang I genannten Technologien bestimmt:

> „a) Konzepte des maschinellen Lernens, mit beaufsichtigtem, unbeaufsichtigtem und bestärkendem Lernen unter Verwendung einer breiten Palette von Methoden, einschließlich des tiefen Lernens (*Deep Learning*);
> b) Logik- und wissensgestützte Konzepte, einschließlich Wissensrepräsentation, induktiver (logischer) Programmierung, Wissensgrundlagen, Inferenz- und Deduktionsmaschinen, (symbolischer) Schlussfolgerungs- und Expertensysteme;
> c) Statistische Ansätze, Bayessche Schätz-, Such- und Optimierungsmethoden[.]“

Eine solche Aufzählung fehlt nunmehr in der aktuellen Version der Verordnung, Artikel 3 Ziffer 1 KI-VO definiert den Begriff des KI-Systems somit **technologieneutral**.[6] Das bedeutet, dass es nicht

1 Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024 zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz und zur Änderung der Verordnungen (EG) Nr. 300/2008, (EU) Nr. 167/2013, (EU) Nr. 168/2013, (EU) 2018/858, (EU) 2018/1139 und (EU) 2019/2144 sowie der Richtlinien 2014/90/EU, (EU) 2016/797 und (EU) 2020/1828 (Verordnung über künstliche Intelligenz) (ABl. L, 2024/1689, 12.7.2024, ELI: http://data.europa.eu/eli/reg/2024/1689/oj).

2 C(2025) 924 final (https://ec.europa.eu/newsroom/dae/redirection/document/112455; Abruf 28.02.2025).

3 Ebd., Absatz 7.

4 Ebd., Absatz 62. Vorschläge zu einer näheren Definition finden sich ferner in *Wendehorst/Nessler/Aufreiter/Aichinger*, Der Begriff des „KI-Systems“ unter der neuen KI-VO, MMR 2024, 605.

5 COM/2021/206 final.

6 Vgl. *Helminger*, Datenschutzrechtliche Herausforderungen bei der Verwendung von Trainingsdaten, EALR 2022, 46, mwN.

mehr so sehr auf technische, informatische Einzelheiten ankommt als vielmehr auf die Relation des Systems zu seiner Umwelt, insbesondere zu menschlichen Personen: Die Wirkung auf das Leben von Menschen steht im Mittelpunkt. Man spricht in diesen Zusammenhängen auch von „**menschenzentrierter** KI“.[7] Dies macht den Begriff offener und damit interpretationsbedürftig, wobei sich die Interpretation in letzter Konsequenz nicht nach theoretischen, sondern nach praktischen Kriterien richtet. Allerdings dürfte es zulässig sein anzunehmen, dass die im Anhang des überholten Entwurfs aufgezählten Technologien in der Regel als KI-Systeme „typischerweise“ der Begriffsbestimmung der aktuellen KI-VO entsprechen. Da sich die Unionslegislative jedoch bewusst dazu entschieden hat, die Aufzählung des alten Anhang I nicht zu übernehmen, darf wiederum der Beweis des Gegenteils nicht ausgeschlossen sein. Dass eine Technologie in dieser Aufzählung enthalten ist, stellt also weder eine hinreichende noch eine notwendige Bedingung für die Anwendbarkeit der KI-VO dar, sondern lediglich eine Art Anhaltspunkt.

Zusätzlich wird zur Eingrenzung des Begriffs der künstlichen Intelligenz im Erwägungsgrund 12 der Verordnung ausgeführt:

> „Der Begriff „KI-System“ in dieser Verordnung sollte klar definiert und eng mit der Tätigkeit internationaler Organisationen abgestimmt werden, die sich mit KI befassen, um Rechtssicherheit, mehr internationale Konvergenz und hohe Akzeptanz sicherzustellen und gleichzeitig Flexibilität zu bieten, um den raschen technologischen Entwicklungen in diesem Bereich Rechnung zu tragen. Darüber hinaus sollte die Begriffsbestimmung auf den wesentlichen Merkmalen der KI beruhen, die sie von einfacheren herkömmlichen Softwaresystemen und Programmierungsansätzen abgrenzen, und sollte sich **nicht** auf Systeme beziehen, die auf **ausschließlich von natürlichen Personen definierten Regeln** für das automatische Ausführen von Operationen beruhen. **Ein wesentliches Merkmal von KI-Systemen ist ihre Fähigkeit, abzuleiten**. Diese Fähigkeit bezieht sich auf den Prozess der Erzeugung von Ausgaben, wie Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen, die physische und digitale Umgebungen beeinflussen können, sowie auf die Fähigkeit von KI-Systemen, Modelle oder Algorithmen oder beides aus Eingaben oder Daten abzuleiten. Zu den Techniken, die während der Gestaltung eines KI-Systems das Ableiten ermöglichen, gehören Ansätze für maschinelles Lernen, wobei aus Daten gelernt wird, wie bestimmte Ziele erreicht werden können, sowie logik- und wissensgestützte Konzepte, wobei aus kodierten Informationen oder symbolischen Darstellungen der zu lösenden Aufgabe abgeleitet wird. Die Fähigkeit ei-

7 Vgl. ErwGr 6 KI-VO: Künstliche Intelligenz „sollte den Menschen als Instrument dienen und letztendlich das menschliche Wohlergehen verbessern.“

nes KI-Systems, abzuleiten, geht über die einfache Datenverarbeitung hinaus, indem Lern-, Schlussfolgerungs- und Modellierungsprozesse ermöglicht werden. Die Bezeichnung „**maschinenbasiert**“ bezieht sich auf die Tatsache, dass KI-Systeme von Maschinen betrieben werden.

Durch die Bezugnahme auf explizite oder implizite **Ziele** wird betont, dass KI-Systeme gemäß explizit festgelegten Zielen oder gemäß impliziten Zielen arbeiten können. Die Ziele des KI-Systems können sich – unter bestimmten Umständen – von der Zweckbestimmung des KI-Systems unterscheiden. Für die Zwecke dieser Verordnung sollten **Umgebungen** als Kontexte verstanden werden, in denen KI-Systeme betrieben werden, während die von einem KI-System erzeugten Ausgaben verschiedene Funktionen von KI-Systemen widerspiegeln, darunter Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen. KI-Systeme sind mit verschiedenen Graden der **Autonomie** ausgestattet, was bedeutet, dass sie bis zu einem gewissen Grad **unabhängig von menschlichem Zutun agieren** und in der Lage sind, ohne menschliches Eingreifen zu arbeiten. Die **Anpassungsfähigkeit**, die ein KI-System nach Inbetriebnahme aufweisen könnte, bezieht sich auf seine **Lernfähigkeit**, durch [d]ie es sich während seiner Verwendung verändern kann. KI-Systeme können eigenständig oder als Bestandteil eines Produkts verwendet werden, unabhängig davon, ob das System physisch in das Produkt integriert (eingebettet) ist oder der Funktion des Produkts dient, ohne darin integriert zu sein (nicht eingebettet).“

Für eine mögliche Vertiefung des Verständnisses des rechtlichen bzw. rechtsrelevanten KI-Begriffs sei auch noch ergänzend die Definition in den Ethik-Leitlinien der Hochrangigen Expertengruppe für künstliche Intelligenz, auf welche die Erwägungsgründe 7 und 27 KI-VO verweisen, angeführt:

„Künstliche-Intelligenz-(KI)-Systeme sind vom Menschen entwickelte Software- (und möglicherweise auch Hardware-) Systeme, die in Bezug auf ein **komplexes Ziel** auf physischer oder digitaler Ebene agieren, indem sie ihre Umgebung durch Datenerfassung **wahrnehmen**, die gesammelten strukturierten oder unstrukturierten Daten **interpretieren**, **Schlussfolgerungen** daraus ziehen oder die aus diesen Daten abgeleiteten Informationen verarbeiten und über die geeignete(n) Maßnahme(n) zur Erreichung des vorgegebenen Ziels **entscheiden**. KI-Systeme können entweder symbolische Regeln verwenden oder ein numerisches Modell erlernen, und sie können auch ihr **Verhalten anpassen**, indem sie analysieren, wie die **Umgebung** von ihren vorherigen Aktionen **beeinflusst** wird.“[8]

8 *Unabhängige Hochrangige Expertengruppe für künstliche Intelligenz*, Ethik-Leitlinien für eine vertrauenswürdige KI (2019), Rz 143; dazu vertiefend *dieselbe*, Eine Definition der KI: Wichtigste Fähigkeiten und Wissensgebiete (2019). Beides unter https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, letzter Abruf

Da Ethik und Recht zwei Gebiete sind, welche in engem Zusammenhang zueinander stehen und beide auf jeweils eigene Weise das „richtige Handeln" zum Gegenstand haben,[9] dürfen generell die Ausführungen der Expertengruppe zur Auslegung der KI-VO herangezogen werden – zumal dies nach deren Erwägungsgründen und gesamter Konzeption nahegelegt und gewollt erscheint.

Abschließend sei zur KI-Definition bemerkt, dass der Begriff der Künstlichen Intelligenz, bzw. des KI-Systems überhaupt, nicht eng, sondern möglichst **weit ausgelegt** werden sollte.[10] Es handelt sich um einen tendenziell unbestimmten und offenen, daher interpretationsbedürftigen **Rechtsbegriff**, nicht um einen exakten Begriff der Informatik oder Naturwissenschaften. Im Allgemeinen darf gesagt werden, dass das Recht als System normal-sprachlicher Regeln oft nicht den Anforderungen an Präzision genügt, die in der Informatik oder Mathematik an operationalisierbare Regeln gestellt werden[11] – wie die Rechtswissenschaft auch nach geisteswissenschaftlichen Methoden verfährt, bei denen es nicht so sehr auf Beobachtung, Formalisierung und Experimente ankommt als vielmehr auf Auslegung und praktische Überlegungen wie beispielsweise die Frage nach der öffentlichen Sicherheit, nach den Grundrechten der betroffenen Individuen, der Transparenz in Bezug auf eine durchschnittlich gebildete Person, dem effektiven Rechtsschutz, … Nicht die Beschreibung natürlicher Vorgänge, sondern die Begründung oder Rechtfertigung von Handlungen – sei es von Privatpersonen oder Staatsorganen – steht im Vordergrund.

In den verschiedenen Wissenschaften, welche sich mit künstlicher Intelligenz beschäftigen, handelt es sich bei dieser vorerst um einen umstrittenen Begriff, das heißt: Es gibt noch keinen genügenden Konsens über die richtige Definition.[12] Einer Definition durch die Informatik, die Kognitions- oder Humanwissenschaften etc. kann die Rechtswissenschaft innerhalb der objektiven Grenzen ihrer eigenen Methoden nicht vorgreifen.[13] Dass der KI-Begriff der Verordnung als solcher weit ausgelegt werden sollte, ergibt sich außerdem daraus, dass in einem zweiten Schritt noch eine engere Auslese oder Einschränkung der fraglichen Systeme vorgenommen wird, nämlich bei der Einstufung in das Risikoschema im Sinne des „**risikobasierten** Ansatzes". Die konkreten **Schutzziele** der KI-VO ergeben sich somit aus zwei Schritten: Zuerst wird ermittelt, ob es sich bei etwas überhaupt um ein

02.07.2024.

9 Vgl. zu diesem Verhältnis *Raz*, The Authority of Law, 233 ff; *Hönigswald*, „Begriff" und „Interesse". Eine methodologische Skizze, AcP 1934, 1.

10 Vgl. *Helminger*, Datenschutzrechtliche Herausforderungen bei der Verwendung von Trainingsdaten, EALR 2022, 46, mwN.

11 Vgl. *Ashley,* Artificial Intelligence and Legal Analytics, 38 ff; *Schweighofer*, Rechtsinformatik und Wissensrepräsentation, 39; *Hart*, The Concept of Law, 124 ff („open texture of law").

12 Vgl. *Pfister*, Does the Criticality Pyramid of the Proposed European AI Act Have Blind Spots? A Comparative Analysis of EU and US Law (https://law.stanford.edu/publications/no-117-does-the-criticality-pyramid-of-the-proposed-european-ai-act-have-blind-spots-a-comparative-analysis-of-eu-and-us-law/, Abruf 05.07.2024), 3 ff.

13 Vgl. *Kelsen*, Reine Rechtslehre, 21 f, gegen „Methodensynkretismus".

System künstlicher Intelligenz handelt, wobei der offenen Definition gemäß kein enger Maßstab angelegt werden sollte – und erst danach wird einschränkend beurteilt, ob die entsprechende Anwendung ein Risiko darstellt, welcher Natur und von welcher Tragweite dieses ist. Auf den Begriff des Risikos in der KI-VO sowie die Einordnung von Anwendungen in einer „Risikopyramide“ wird im nächsten Abschnitt eingegangen.

2. Der „risikobasierte Ansatz“; Begriff des Risikos; „Schaden“; Einstufung von KI

Der Begriff des Risikos wird als ein „Herzstück“ der KI-VO betrachtet. Dementsprechend wird im Erwägungsgrund 26 die Konzeption der Verordnung als „risikobasierter Ansatz“ bezeichnet:

> „Um ein verhältnismäßiges und wirksames verbindliches Regelwerk für KI-Systeme einzuführen, sollte ein klar definierter risikobasierter Ansatz verfolgt werden. Bei diesem Ansatz sollten Art und Inhalt solcher Vorschriften **auf die Intensität und den Umfang der Risiken zugeschnitten** werden, die von KI-Systemen ausgehen können. Es ist daher notwendig, bestimmte **inakzeptable Praktiken im Bereich der KI zu verbieten** und Anforderungen an **Hochrisiko-KI-Systeme** und Pflichten für die betreffenden Akteure sowie Transparenzpflichten für bestimmte KI-Systeme festzulegen.“

Was ist überhaupt ein Risiko im normativen Sinn der Verordnung? Diese Frage wird in Artikel 3 Ziffer 2 auf einfache und eingängige Weise durch den Begriff des **Schadens** beantwortet:

> „'Risiko' [ist] die Kombination aus der **Wahrscheinlichkeit** des Auftretens eines Schadens und der **Schwere** dieses Schadens[.]“

Es gibt im österreichischen Zivilrecht eine eigene Definition des Schadens, welche in Auslegung des Allgemeinen Bürgerlichen Gesetzbuches von Wissenschaft und Rechtsprechung über lange Zeit sehr ausführlich und differenziert entwickelt wurde. Allerdings handelt es sich bei der KI-VO um Unionsrecht, welches unabhängig von nationalem Recht ist und auch unabhängig davon ausgelegt werden sollte: Man spricht in diesem Zusammenhang von „**autonomer Interpretation**“. Es soll in der ganzen Europäischen Union für den gesamten Anwendungsbereich der KI-VO einen einheitlichen und möglichst unkomplizierten Schadensbegriff geben, der nicht von den – teils sehr unterschiedlichen – historischen Rechtstraditionen der einzelnen Mitgliedstaaten belastet ist. Das hängt damit zusammen, dass es in der EU einen „Binnenmarkt“ geben soll, in dem Produkte wie beispielsweise KI ohne zwischenstaatliche Hindernisse wie unter anderem divergierende Rechtsbegrif-

fe, -vorschriften oder -anschauungen zirkulieren können. Bereits der erste Erwägungsgrund der Verordnung bezeichnet den europäischen Binnenmarkt als „den“ Zweck schlechthin für die Regulierung von KI:

> „Zweck dieser Verordnung ist es, das **Funktionieren des Binnenmarkts** zu verbessern, indem ein **einheitlicher Rechtsrahmen** insbesondere für die Entwicklung, das Inverkehrbringen, die Inbetriebnahme und die Verwendung von Systemen künstlicher Intelligenz (KI-Systeme) in der Union im Einklang mit den Werten der Union festgelegt wird, um die Einführung von menschenzentrierter und vertrauenswürdiger künstlicher Intelligenz (KI) zu fördern und **gleichzeitig ein hohes Schutzniveau** in Bezug auf Gesundheit, Sicherheit und der in der Charta der Grundrechte der Europäischen Union („Charta“) verankerten Grundrechte, einschließlich Demokratie, Rechtsstaatlichkeit und Umweltschutz, sicherzustellen, den Schutz vor schädlichen Auswirkungen von KI-Systemen in der Union zu gewährleisten und gleichzeitig die Innovation zu unterstützen. Diese Verordnung gewährleistet den grenzüberschreitenden freien Verkehr KI-gestützter Waren und Dienstleistungen, wodurch verhindert wird, dass die Mitgliedstaaten die Entwicklung, Vermarktung und Verwendung von KI-Systemen beschränken, sofern dies nicht ausdrücklich durch diese Verordnung erlaubt wird.“[14]

Wie aus dieser Textpassage hervorgeht, ist auch für ein „hohes Schutzniveau“ der Sicherheit und Grundrechte ein einheitliches, sich nicht je nach Mitgliedstaat und Rechtskultur unterscheidendes Verständnis des Unionsrechts notwendig. Das bedeutet nicht, dass die vielfältigen historischen Traditionen als solche irrelevant wären – vielmehr besteht die Aufgabe und Herausforderung darin, eine gemeinsame Grundlage von Werten zu ermitteln, die in ganz Europa bzw. in der ganzen EU wichtig sind. Zudem verfolgt die Verordnung zusätzliche, neuartige regulatorische Ziele und erhebt den Anspruch, maßgeblich zum technologischen wie gesellschaftlichen Fortschritt beizutragen und gerade dabei den einzelnen mitgliedstaatlichen Rechtssystemen voraus zu sein. Aus dem allem folgt, dass den Begriffen der KI-VO ein eigenes, unionsrechtliches Verständnis zugrundezulegen ist, welches vom nationalen Recht unabhängig ist. Dabei scheint es einen weiteren Spielraum für intuitive Wertungen zu geben, weil in der EU grundlegende Begriffe wie zum Beispiel der des „Schadens“ vergleichsweise wenig im Detail festgelegt, sondern eher auf allgemeine, grundsätzliche Weise bestimmt sind.[15] Das legt es nahe, dass man auch ohne juristische Ausbildung oder vertiefendes rechtliches Vorwissen sinnvoll Überlegungen zum Schadensbegriff entwickeln kann. Zumal die KI-VO gibt ausdrücklich zu erkennen, dass unter einem „Schaden“ im Rechtssinn sehr vielfältige und

14 Dasselbe sagt in kürzerer Form Artikel 1 KI-VO.

15 Vgl. *Wurmnest*, Grundzüge eines europäischen Haftungsrechts (2003), 194 ff; *Oskierski*, Schadenersatz im Europäischen Recht. Eine vergleichende Untersuchung des Acquis Communitaire und der EMRK (2010).

verschiedenartige Nachteile verstanden werden können, ohne diejenigen Einschränkungen, die in nationalen Rechtssystemen wie insbesondere in Österreich oftmals für rechtsrelevante Schäden gemacht werden:[16]

> „KI [kann] je nach den Umständen ihrer konkreten Anwendung und Nutzung sowie der technologischen Entwicklungsstufe Risiken mit sich bringen und öffentliche Interessen und grundlegende Rechte schädigen, die durch das Unionsrecht geschützt sind. Ein solcher Schaden kann materieller oder immaterieller Art sein, einschließlich **physischer, psychischer, gesellschaftlicher oder wirtschaftlicher** Schäden." (ErwGr 5)

Welcher Schaden erwächst beispielsweise dem Verdächtigen einer Straftat, wenn Polizei- und Justizbehörden gegen ihn mithilfe von KI-Anwendungen ermitteln und dabei etwas „schiefläuft"? Beispielsweise könnte die KI aufgrund eines Vorurteils die ermittelnden Organe zu Entscheidungen bewegen, welche tiefgreifend in das Leben der betreffenden Person eingreifen, beispielsweise Hausdurchsuchung, Festnahme, Untersuchungshaft, … Die Situation, einer Strafverfolgung ausgesetzt zu sein, bedeutet als solche einen spürbaren psychischen Schaden im Sinne einer schweren Beeinträchtigung des Alltags und Wohlbefindens. Das seelische Leid kann zudem zu Schäden an der physischen Gesundheit führen, die wiederum mithilfe finanzieller Aufwendungen behandelt oder gelindert werden müssen, wodurch schließlich auch wirtschaftliche Schäden entstehen. Nicht zuletzt bedeutet ein Strafverfahren gegen eine sonst als achtbar angesehene Person eine erhebliche Stigmatisierung,[17] wenn andere Menschen davon erfahren, beispielsweise die Familie, das berufliche Umfeld, der Freundes- oder Bekanntenkreis; in besonderen Fällen fangen auch die Medien an, sich für Strafverfahren zu interessieren und dabei bestimmte Vermutungen anzustellen. Das alles führt zu einem schweren gesellschaftlichen Schaden, da der angeschuldigten Person ihr guter Ruf, ihre soziale Akzeptiertheit, ihre unbefangene Lebensführung im Kontext anderer Menschen weggenommen werden und sie in bestimmten Bereichen ausgeschlossen wird, nicht mehr willkommen ist. Freilich entsteht ferner beim Verlust des Arbeitsplatzes ein wirtschaftlicher Schaden durch den Einkommensverlust. Welcher Natur wäre dann ein Schaden, wenn eine Person durch einen solchen Druck keine andere Option sieht, als Suizid zu begehen? Wie ist mit den Schäden umzugehen, die in weiterer Folge bei jenen Personen entstehen, welche dem Opfer einer unrichtigen KI-Anwendung nahestanden, wie ist mit Trauerzuständen umzugehen oder mit dem Verlust einer Person, die für andere

16 Siehe zB *Palmer*, European contractual regimes: The contemporary approaches. In: *Palmer* (ed.), The Recovery of Non-Pecuniary Loss in European Contract Law (2015), 95 [108 ff], wo es generell um den besonders eingeschränkten Ersatz immaterieller Schäden in Österreich geht.

17 Vgl. für viele *Zerbes*, Spitzeln, Spähen, Spionieren. Sprengung strafprozessualer Grenzen durch geheime Zugriffe auf Kommunikation (2011), 268.

vormals finanziell aufgekommen war, oder mit dem Verlust einer geliebten Person, die einem unersetzlich ist? All diese Möglichkeiten zeigen auf, dass durch eine problematische Verwendung von KI in einem so sensiblen Bereich wie der Strafverfolgung rasch vielfältige wie schwerwiegende Schäden an Personen und deren Vermögen auftreten können. – Allen diesen Schäden möchte die KI-VO zuvorkommen, indem sie den zentralen Begriff des Risikos nach ihnen ausrichtet und hochriskante KI-Systeme strengen Vorschriften unterwirft.

Im vorliegenden Projekt interessiert vor allem der Einsatz von Künstlicher Intelligenz in der „**Strafverfolgung**", welche in Artikel 3 Ziffer 46 KI-VO eigens definiert ist als „Tätigkeiten der Strafverfolgungsbehörden oder in deren Auftrag zur **Verhütung, Ermittlung, Aufdeckung oder Verfolgung** von Straftaten oder zur Strafvollstreckung, **einschließlich des Schutzes vor und der Abwehr von Gefahren für die öffentliche Sicherheit**". Es handelt sich hier um ein aus österreichischer Sicht ungewöhnliches Begriffsverständnis, weil im nationalen Recht bisher strikt zwischen der Strafverfolgung im engeren Sinne nach der Strafprozessordnung, welche sich auf der Ermittlung bereits begangener Taten bezieht, und der polizeilichen Gefahrenabwehr (bzw. „Verhütung") nach dem Sicherheitspolizeigesetz, welche zukünftiges Fehlverhalten zum Gegenstand hat, unterschieden wurde.[18] Erstere wird im System der Gewaltenteilung als Aufgabe der Justiz, in deren Dienste die Kriminalpolizei arbeitet, zweitere als Aufgabe der Verwaltung, der Sicherheitspolizei gesehen – diese Aufgaben unterliegen jeweils nicht nur unterschiedlichen Normen, sondern beinhalten auch eine unterschiedliche Weisungsstruktur: Während die Kriminalpolizei funktional der Staatsanwaltschaft untersteht, richtet sich die Sicherheitspolizei nach dem Innenministerium.

Nun wird auch die „**Strafverfolgungsbehörde**" in Art 3 Z 45 KI-VO in einem Atemzug ohne Rücksicht auf Unterschiede zwischen Justiz und Sicherheitspolizei folgendermaßen definiert:

> „a) eine Behörde, die für die Verhütung, Ermittlung, Aufdeckung oder Verfolgung von Straftaten oder die Strafvollstreckung, einschließlich des Schutzes vor und der Abwehr von Gefahren für die öffentliche Sicherheit, zuständig ist, oder
> b) eine andere Stelle oder Einrichtung, der durch nationales Recht die Ausübung öffentlicher Gewalt und hoheitlicher Befugnisse zur Verhütung, Ermittlung, Aufdeckung oder Verfolgung von Straftaten oder zur Strafvollstreckung, einschließlich des Schutzes vor und der Abwehr von Gefahren für die öffentliche Sicherheit, übertragen wurde[.]"

18 Vgl. ebd., Kapitel IV.

Das bedeutet, dass „Strafverfolgungsbehörden“ im Sinne der KI-VO nicht nur Justizbehörden wie Gerichte, Staatsanwaltschaften und – funktionell an diese gebunden – Kriminalpolizei, sondern auch Sicherheitsbehörden sind.

Nun unterscheidet die KI-VO in einer Art „Risikopyramide“ verschiedene Arten von KI-Systemen, je nach Art und Ausmaß von Risiken: Zum einen gibt es „**verbotene Praktiken** im KI-Bereich“ gemäß Art 5 KI-VO wie beispielsweise diskriminierende, manipulative Anwendungen, *social scoring*, *predictive policing*,[19] das ungezielte Auslesen von Gesichtsbildern aus dem Internet oder von Überwachungsaufnahmen sowie die Verwendung biometrischer Echtzeit-Fernidentifizierungssysteme in öffentlich zugänglichen Räumen zu Strafverfolgungszwecken, welche letztere jedoch wiederum für bestimmte besonders dringende Zwecke unter Art 5 Abs 1 lit h ausnahmsweise erlaubt wird;[20] sodann werden in Art 6 „**Hochrisiko-KI-Systeme**“ derart definiert, dass einerseits Produkte nach den im Anhang I aufgezählten EU-Produktsicherheitsvorschriften und andererseits die im Anhang III[21] genannten Systeme darunter verstanden werden, insbesondere Biometrie, kritische Infrastruktur, Strafverfolgung, Migration samt Asyl sowie Grenzkontrolle und Rechtspflege; schließlich kennt die Verordnung Systeme mit mittlerem („Transparenzrisiko“) und solche mit minimalem Risiko,[22] die im vorliegenden Zusammenhang genauso wie die Kategorie der „KI-Modelle mit allgemeinem Verwendungszweck“ jedoch nicht behandelt werden müssen. Überhaupt **ausgenommen** vom Anwendungsbereich der KI-VO sind unter anderem gemäß Art 2 Abs 6 solche KI-Systeme, die nur für die wissenschaftliche Forschung oder Entwicklung verwendet werden, allerdings gleichermaßen auch die besonders sicherheits- und grundrechtskritischen Systeme für militärische Zwecke, Verteidigungszwecke oder Zwecke der nationalen Sicherheit gemäß Abs 3.

Zudem werden die hochriskanten KI-Systeme der Strafverfolgung unter dem Vorbehalt, dass sie nach nationalem oder EU-Recht zulässig seien, im Anhang III der Verordnung unter Ziffer 6 näher umschrieben:

> „a) KI-Systeme, die bestimmungsgemäß von Strafverfolgungsbehörden oder in deren Namen oder von Organen, Einrichtungen und sonstigen Stellen der Union zur Unterstützung

19 Nach Art 5 Abs 1 lit d jedoch mit Ausnahme derjenigen KI-Systeme, „die dazu verwendet werden, die durch Menschen durchgeführte Bewertung der Beteiligung einer Person an einer kriminellen Aktivität, die sich bereits auf objektive und überprüfbare Tatsachen stützt, die in unmittelbarem Zusammenhang mit einer kriminellen Aktivität stehen, zu unterstützen“.

20 Dazu näher unten.

21 Gemäß Art 7 KI-VO kann die Europäische Kommission unter bestimmten Bedingungen Änderungen an Anhang III vornehmen.

22 Vgl. *Pollirer*, Checkliste AI Act, Dako 2024, 62.

von Strafverfolgungsbehörden oder in deren Namen zur Bewertung des **Risikos einer natürlichen Person, zum Opfer von Straftaten zu werden**, verwendet werden sollen;

b) KI-Systeme, die bestimmungsgemäß von Strafverfolgungsbehörden oder in deren Namen oder von Organen, Einrichtungen und sonstigen Stellen der Union zur Unterstützung von Strafverfolgungsbehörden als **Lügendetektoren** oder ähnliche Instrumente verwendet werden sollen;

c) KI-Systeme, die bestimmungsgemäß von Strafverfolgungsbehörden oder in deren Namen oder von Organen, Einrichtungen und sonstigen Stellen der Union zur Unterstützung von Strafverfolgungsbehörden zur Bewertung der **Verlässlichkeit von Beweismitteln** im Zuge der Ermittlung oder Verfolgung von Straftaten verwendet werden sollen;

d) KI-Systeme, die bestimmungsgemäß von Strafverfolgungsbehörden oder in deren Namen oder von Organen, Einrichtungen und sonstigen Stellen der Union zur Unterstützung von Strafverfolgungsbehörden zur Bewertung des **Risikos, dass eine natürliche Person eine Straftat begeht oder erneut begeht**, nicht nur auf der Grundlage der Erstellung von Profilen natürlicher Personen gemäß Artikel 3 Absatz 4 der Richtlinie (EU) 2016/680 oder zur **Bewertung persönlicher Merkmale und Eigenschaften oder vergangenen kriminellen Verhaltens** von natürlichen Personen oder Gruppen verwendet werden sollen;

e) KI-Systeme, die bestimmungsgemäß von Strafverfolgungsbehörden oder in deren Namen oder von Organen, Einrichtungen und sonstigen Stellen der Union zur Unterstützung von Strafverfolgungsbehörden zur Erstellung von **Profilen natürlicher Personen** gemäß Artikel 3 Absatz 4 der Richtlinie (EU) 2016/680 im Zuge der Aufdeckung, Ermittlung oder Verfolgung von Straftaten verwendet werden sollen."

Darüber hinaus zählen **biometrische** KI-Systeme – ob nun im Bereich Strafverfolgung oder in einem anderen verwendet – gemäß Anhang III Z 1 immer als Hochrisikosysteme. Sie werden darin folgendermaßen definiert:

„a) biometrische **Fernidentifizierungssysteme**. Dazu gehören nicht KI-Systeme, die bestimmungsgemäß für die biometrische Verifizierung, deren einziger Zweck darin besteht, zu bestätigen, dass eine bestimmte natürliche Person die Person ist, für die sie sich ausgibt, verwendet werden sollen;

b) KI-Systeme, die bestimmungsgemäß für die biometrische **Kategorisierung** nach sensiblen oder geschützten Attributen oder Merkmalen auf der Grundlage von Rückschlüssen auf diese Attribute oder Merkmale verwendet werden sollen;

c) KI-Systeme, die bestimmungsgemäß zur **Emotionserkennung** verwendet werden sollen."

Es sei in diesem Zusammenhang noch bemerkt, dass der Begriff *predictive policing* gewissermaßen ein Grenzgebiet zwischen einer verbotenen KI-Praktik und einem streng regulierten, aber an sich zulässigen Hochrisiko-KI-System bildet. Dies geht schon im obigen Zitat aus lit d hervor, wo die „Bewertung des Risikos, dass eine natürliche Person eine Straftat begeht oder erneut begeht", als eine mögliche legale Vorgehensweise angesprochen wird. Ferner verbietet Art 5 Abs 1 lit d KI-VO „das Inverkehrbringen, die Inbetriebnahme für diesen spezifischen Zweck oder die Verwendung eines KI-Systems zur Durchführung von Risikobewertungen in Bezug auf natürliche Personen, um das Risiko, dass eine natürliche Person eine Straftat begeht, ausschließlich auf der Grundlage des Profiling einer natürlichen Person oder der Bewertung ihrer persönlichen Merkmale und Eigenschaften zu bewerten oder vorherzusagen" – dieses Verbot gilt jedoch wohlgemerkt nicht für KI-Systeme, „die dazu verwendet werden, die durch Menschen durchgeführte Bewertung der Beteiligung einer Person an einer kriminellen Aktivität, die sich bereits auf objektive und überprüfbare Tatsachen stützt, die in unmittelbarem Zusammenhang mit einer kriminellen Aktivität stehen, zu **unterstützen**". Man kann also unterscheiden zwischen „rein" bzw. „ausschließlich" KI-gestütztem *predictive policing* und einem solchen, das lediglich die menschliche Ermittlungsarbeit unterstützt.[23] Ausschließlich letzteres ist nach der KI-VO zulässig: Die Ergebnisse der KI-Anwendung dürfen bei der behördlichen Aufgabenerfüllung nicht ausschlaggebend sein bzw. unkritisch hingenommen und als Grundlage für Grundrechtseingriffe genutzt werden,[24] sondern es muss stets bei der menschliches Denken oder Schlussfolgern bloß begleitenden Nutzung bleiben. Dabei wird sich auch die Frage stellen, ob die betreffende Anwendung nur in der sicherheitspolizeilichen Gefahrenabwehr eine Rolle spielen wird oder auch in der richterlichen Arbeit der Strafbemessung, bei welcher die Wahrscheinlichkeit, dass ein strafbares Verhalten wiederholt werden könnte, eine gewisse Bedeutung hat.[25]

Nicht jedes KI-System, welches im Bereich der Strafverfolgung eingesetzt wird, ist automatisch ein Hochrisiko-System. Art 6 Abs 3 KI-VO sieht für bestimmte, eher nebensächliche Tätigkeiten eine **Ausnahme** vor:

> „Abweichend von Absatz 2 gilt ein in Anhang III genanntes KI-System nicht als hochriskant, wenn es **kein erhebliches Risiko** der Beeinträchtigung in Bezug auf die Gesundheit,

23 Vgl. auch die Beispiele im Leitlinienentwurf der Kommission gemäß Art 96 KI-VO über verbotene Praktiken: C(2025) 884 final (https://ec.europa.eu/newsroom/dae/redirection/document/112367; Abruf 28.02.2025), Absatz 214. Zuvor schon EuGH 21.06.2022, C-817/19 (*Ligue des droits humains*).

24 Siehe unten zum Thema „menschliche Aufsicht".

25 Während KI-Anwendungen im strafprozessualen bzw. sicherheitspolizeilichen Ermittlungsverfahren dem Bereich „Strafverfolgung" nach Anhang III Z 6 KI-VO unterfallen, gehören solche, die mit der gerichtlichen Strafbemessung im Zusammenhang stehen, zum Bereich „Rechtspflege und demokratische Prozesse" nach Z 8.

Sicherheit oder Grundrechte natürlicher Personen birgt, indem es unter anderem **nicht das Ergebnis der Entscheidungsfindung wesentlich beeinflusst**.

Unterabsatz 1 gilt, wenn eine der folgenden Bedingungen erfüllt ist:

a) das KI-System ist dazu bestimmt, eine eng gefasste Verfahrensaufgabe durchzuführen;

b) das KI-System ist dazu bestimmt, das Ergebnis einer zuvor abgeschlossenen menschlichen Tätigkeit zu verbessern;

c) das KI-System ist dazu bestimmt, Entscheidungsmuster oder Abweichungen von früheren Entscheidungsmustern zu erkennen, und ist nicht dazu gedacht, die zuvor abgeschlossene menschliche Bewertung ohne eine angemessene menschliche Überprüfung zu ersetzen oder zu beeinflussen; oder

d) das KI-System ist dazu bestimmt, eine vorbereitende Aufgabe für eine Bewertung durchzuführen, die für die Zwecke der in Anhang III aufgeführten Anwendungsfälle relevant ist.

Ungeachtet des Unterabsatzes 1 gilt ein in Anhang III aufgeführtes KI-System immer dann als hochriskant, wenn es ein **Profiling** natürlicher Personen vornimmt."[26]

Angesichts des letzten Satzes dürfte der Anwendungsbereich dieser Ausnahme im Bereich der Strafverfolgung relativ gering sein: Denn wenn von der Behörde mithilfe von KI Ermittlungen geführt werden – das heißt: Ermittlungen gegen natürliche Personen oder Gruppen derselben –, geht es immer auch um die Erstellung von Profilen über natürliche Personen, insbesondere im Rahmen eines Aktes zu einer möglicherweise bestehenden Gefahr oder zu einer mutmaßlich begangenen Straftat, damit verbunden zu einer oder mehreren Personen, die damit im Zusammenhang stehen und über welche Informationen gesammelt, analysiert, verknüpft und zusammengefasst werden. So wird in § 36 Abs 2 Z 4 DSG „Profiling" definiert als „jede Art der automatisierten Verarbeitung personenbezogener Daten, die darin besteht, dass diese personenbezogenen Daten verwendet werden, um bestimmte persönliche Aspekte, die sich auf eine natürliche Person beziehen, zu bewerten, insbesondere um Aspekte bezüglich Arbeitsleistung, wirtschaftliche Lage, Gesundheit, persönliche Vorlieben, Interessen, Zuverlässigkeit, Verhalten, Aufenthaltsort oder Ortswechsel dieser natürlichen Person zu analysieren oder vorherzusagen".[27] Laut ErwGr 53 sollen beispielsweise KI-Anwendungen, deren Funktion ausschließlich in der **Übersetzung** von Texten aus einer natürlichen Sprache in eine andere besteht, keine Hochrisikosysteme sein; ferner solche Programme, welche der Verbesserung

26 In **prozeduraler Hinsicht** wird dabei Folgendes bestimmt: Stuft der Anbieter sein System in diesem Kontext als nicht hochriskant ein, hat er dies vor dem Inverkehrbringen oder Inbetriebnehmen gemäß Art 6 Abs 4 KI-VO zu dokumentieren und das System nach Art 49 Abs 2 in der EU-Datenbank zu registrieren. Die MÜB hat in Bezug auf ein solches KI-System Befugnisse im Sinne des Art 80.
Für die praktische Anwendung der Ausnahmen nach Art 6 Abs 3 wird die Europäische Kommission laut Abs 5 **Leitlinien** erarbeiten.

27 Vgl. Art 3 Z 52 KI-VO, mit Verweis auf die gleichlautende Definition in der DSGVO, sowie Art 3 Z 4 DSRL-PJ.

sprachlicher Formulierungen dienen;[28] sofern in keiner Hinsicht „Profiling“ involviert ist, sind auch klassische **rechtsinformatische Aufgaben**[29] wie die formale oder semantische Analyse von Rechtstexten oder bereits ermittelten Sachverhaltselementen[30] von dieser Ausnahme erfasst, als „intelligente Lösungen für die Bearbeitung von Dossiers, wozu verschiedene Funktionen wie Indexierung, Suche, Text- und Sprachverarbeitung oder Verknüpfung von Daten mit anderen Datenquellen gehören“ (ErwGr 53). Ob das sogenannte „**Wissensmanagement**“ zur Hochrisiko-Kategorie gehört, hängt gleichermaßen davon ab, ob personenbezogene Daten bzw. Daten zu einem konkreten Fall verarbeitet werden: Handelt es sich bloß um **anonyme kriminalstatistische Daten**,[31] beispielsweise im Rahmen einer generellen Analyse und Bewertung der Gefährlichkeit eines Ortes als solchen, ist wohl nicht davon auszugehen, dass es sich dabei um ein Hochrisiko-KI-System handelt.

3. Rollenverteilung und Pflichten nach den Vorschriften der KI-VO über Hochrisiko-Systeme

Kapitel III Abschnitt 2 und 3 der KI-VO regelt auf relativ detaillierte und vielfältige Art die verschiedenen Pflichten, welche die jeweiligen Akteure mit Bezug zu Hochrisiko-KI-Systemen treffen. Die zentralen Rollen sind gemäß Art 3 die des „Anbieters“ und die des „Betreibers“, daneben bestehen „Bevollmächtigte“, „Einführer“ und „Händler“.

Art 3 Z 4 KI-VO definiert den **Betreiber** eines KI-Systems als „eine natürliche oder juristische Person, Behörde, Einrichtung oder sonstige Stelle, die ein KI-System in eigener Verantwortung verwendet, es sei denn, das KI-System wird im Rahmen einer persönlichen und nicht beruflichen Tätigkeit verwendet“. Somit zählen beispielsweise die **Landespolizeidirektion** oder die **Direktion Staatsschutz und Nachrichtendienst** (bzw. die Generaldirektion für die öffentliche Sicherheit) nach dem SNG, wenn sie KI-Systeme im Bereich „Strafverfolgung“ verwenden, als Betreiber im Sinne der KI-VO.

Dagegen ist gemäß Art 3 Z 3 KI-VO ein **Anbieter** eines KI-Systems „eine natürliche oder juristische Person, Behörde, Einrichtung oder sonstige Stelle, die ein KI-System oder ein KI-Modell mit

28 „Diese Bedingung würde beispielsweise für KI-Systeme gelten, deren Ziel es ist, die in zuvor verfassten Dokumenten verwendete Sprache zu verbessern, etwa den professionellen Ton, den wissenschaftlichen Sprachstil oder um den Text an einen bestimmten mit einer Marke verbundenen Stil anzupassen.“ (ErwGr 53 KI-VO)

29 Dazu *Ashley,* Artificial Intelligence and Legal Analytics; *Schweighofer*, Rechtsinformatik und Wissensrepräsentation; und etliche andere.

30 Vgl. *Mayrhofer*, Rechtliche Herausforderungen und Grenzen des Einsatzes von Assistenzsystemen in der Gerichtsbarkeit, RZ 2024, 27 [33] („maschinelle Unterstützung bei der Strukturierung des von einer natürlichen Person ermittelten Sachverhalts“).

31 Die KI-VO selbst enthält allerdings keine speziellen Bestimmungen zum Thema „Statistik“ oder „anonyme Daten“, sodass mit Ausnahmen vorsichtig umzugehen ist.

allgemeinem Verwendungszweck **entwickelt oder entwickeln lässt** und es unter ihrem eigenen Namen oder ihrer Handelsmarke **in Verkehr bringt** oder das KI-System unter ihrem eigenen Namen oder ihrer Handelsmarke **in Betrieb nimmt**, sei es entgeltlich oder unentgeltlich". Unter den Begriff des Anbieters fallen also vor allem **Technologiefirmen**, welche KI programmieren bzw. entwickeln und dem Markt zur Verfügung stellen. Denkbar ist es aber auch, dass eine staatliche Stelle selbst mit dem Programmieren oder Herstellen eines KI-Systems betraut ist; falls sie es dann auch selbst eigenverantwortlich für eine Tätigkeit in einem von der KI-VO umfassten Risikobereich einsetzt, ist sie zugleich Anbieter und Betreiber, insoweit fallen die beiden Rollen in einem einzigen Akteur zusammen. Teils erübrigen sich dadurch jene Pflichten, welche nur für das Verhältnis zwischen Anbieter und Betreiber als zwei voneinander verschiedenen Akteure gelten – beispielsweise die Bereitstellung von Betriebsanleitungen im Sinne des Art 3 Z 15 und Art 13 KI-VO –, teils gelten jedoch für dieselbe Stelle sowohl die anbieter- als auch die betreiberspezifischen Vorschriften,[32] beispielsweise in der Rolle des Betreibers die Grundrechte-Folgenabschätzung gemäß Art 27 KI-VO und in der Rolle des Anbieters das Qualitätsmanagementsystem gemäß Art 17 KI-VO.

Somit sind die beiden grundlegenden Rollen, Anbieter und Betreiber, voneinander einigermaßen klar abgegrenzt. Jedoch kann in diesem Zusammenhang auch eine Art Vermengung oder Verschiebung dieser Zuständigkeiten auftreten, was in Art 25 KI-VO geregelt ist. Danach **gelten Händler, Einführer, Betreiber oder sonstige Dritte als Anbieter** und unterliegen den Anbieterpflichten gemäß Art 16:

> „a) wenn sie ein bereits in Verkehr gebrachtes oder in Betrieb genommenes Hochrisiko-KI-System mit ihrem **Namen** oder ihrer **Handelsmarke** versehen, unbeschadet vertraglicher Vereinbarungen, die eine andere Aufteilung der Pflichten vorsehen;
> b) wenn sie eine **wesentliche Veränderung** eines Hochrisiko-KI-Systems, das bereits in Verkehr gebracht oder in Betrieb genommen wurde, so vornehmen, dass es weiterhin ein Hochrisiko-KI-System gemäß Artikel 6 bleibt;
> c) wenn sie die **Zweckbestimmung** eines KI-Systems, einschließlich eines KI-Systems mit allgemeinem Verwendungszweck, das nicht als hochriskant eingestuft wurde und bereits in Verkehr gebracht oder in Betrieb genommen wurde, so verändern, dass das betreffende KI-System zu einem Hochrisiko-KI-System im Sinne von Artikel 6 wird."

Es ist also eine gewisse Vorsicht geboten, wenn beispielsweise eine Polizeibehörde das in der „Strafverfolgung" oder einem sonstigen Bereich eingesetzte und dabei von einem Anbieter oder

32 So auch ErwGr 83 KI-VO: „In bestimmten Situationen könnten diese Akteure mehr als eine Rolle gleichzeitig wahrnehmen und sollten daher alle einschlägigen Pflichten, die mit diesen Rollen verbunden sind, kumulativ erfüllen."

Händler erworbene KI-System für allgemeine oder auch nur punktuelle Bedürfnisse „weiterentwickelt“ oder in irgendeiner Weise modifiziert bzw. umprogrammiert. In einem solchen Fall kann es passieren, dass eine solche Handlung in einem Verfahren vor der Marktüberwachungsbehörde oder auch vor einem ordentlichen Gericht wegen Amtshaftung als „wesentliche Veränderung“ gewertet wird;[33] dies hätte dann zur Folge, dass die Polizeibehörde für die Erfüllung aller derjenigen Pflichten Verantwortung übernehmen muss, die sonst nur eine große Technologiefirma als ursprünglicher Anbieter trägt bzw. aufgrund der technologischen Spezialisierung allein tragen kann. Daher muss unbedingt, bevor auch nur eine geringe Änderung am KI-System vorgenommen wird, Rücksprache mit dem Anbieter und mit der Marktüberwachungsbehörde bzw. mit der Konformitätsbewertungsstelle gehalten werden, um ein objektiv begründetes Einvernehmen darüber zu erzielen, was als eine „wesentliche Veränderung“ gemäß Art 25 KI-VO zu gelten hat.[34]

Ferner kann durch die „Änderung der Zweckbestimmung“ des KI-Systems im Sinne des Art 25 Abs 1 lit b KI-VO ein gewöhnlicher Betreiber zum Anbieter eines Hochrisiko-KI-Systems werden und dessen Pflichten unterfallen. Als „**Zweckbestimmung**“ definiert Art 3 Z 12 KI-VO „die Verwendung, für die ein KI-System **laut Anbieter** bestimmt ist, einschließlich der **besonderen Umstände und Bedingungen** für die Verwendung, entsprechend den vom Anbieter bereitgestellten Informationen in den **Betriebsanleitungen**, im Werbe- oder Verkaufsmaterial und in diesbezüglichen Erklärungen sowie in der technischen Dokumentation“. Als Beispiel für eine solche Änderung der Zweckbestimmung lässt sich beispielsweise ein – möglicherweise auch biometrisches – KI-System anführen, das ursprünglich allein dem Zweck der **Cybersicherheit**, also insbesondere dem Schutz der IT vor Angriffen durch Dritte, oder des Datenschutzes dient. Ein solches System soll, falls es ausschließlich diesen eng gefassten Zwecken dient, gemäß ErwGr 54 KI-VO an sich nicht als Hochrisiko-KI-System gelten und damit nicht den strengen Vorschriften des Kapitels III unterliegen.[35] Wird jedoch ein solches biometrisches KI-System über diese Zweckbestimmung hinaus von der Polizeibehörde im Bereich „Strafverfolgung“ im Sinne des Art 6 und des Anhangs III KI-VO verwendet, gilt diese Polizeibehörde plötzlich als Anbieter eines Hochrisiko-KI-Systems und muss dessen sämtliche Sorgfaltsmaßstäbe und Aufgaben im Sinne der Verordnung verantworten.

33 Siehe auch Art 3 Z 23 KI-VO, welcher eine wesentliche Veränderung definiert als „eine Veränderung eines KI-Systems nach dessen Inverkehrbringen oder Inbetriebnahme, die in der vom Anbieter durchgeführten ursprünglichen Konformitätsbewertung nicht vorgesehen oder geplant war und durch die die Konformität des KI-Systems mit den Anforderungen in Kapitel III Abschnitt 2 beeinträchtigt wird oder die zu einer Änderung der Zweckbestimmung führt, für die das KI-System bewertet wurde“.

34 Vgl. Art 43 Abs 1, 2 und 4 KI-VO.

35 Vgl. auch die Ausnahme in Anhang III Z 1 lit a KI-VO für „KI-Systeme, die bestimmungsgemäß für die biometrische Verifizierung, deren einziger Zweck darin besteht, zu bestätigen, dass eine bestimmte natürliche Person die Person ist, für die sie sich ausgibt, verwendet werden sollen“.

Ob das bloße „**Trainieren**“ eines KI-Systems eine „wesentliche Veränderung“ oder eine Änderung in der Zweckbestimmung begründet, hängt von der konkreten Situation ab. ErwGr 88 KI-VO legt es nahe, dass das Trainieren als solches allein noch nicht zur Anbietereigenschaft führt.[36] Wenn der Betreiber das System lediglich unter denjenigen Bedingungen und in demjenigen Ausmaß trainiert, welche der Anbieter im Rahmen eines korrekt durchgeführten Risikomanagements (Art 9 KI-VO) in der Betriebsanleitung für das risikoarme selbstverantwortliche Trainieren durch den Betreiber vorgesehen hat,[37] ist davon auszugehen, dass dieser dadurch seine Rolle nicht wechselt. Es liegt in der Verantwortung des Anbieters, einen vernünftigen Rahmen für das Trainieren durch Betreiber festzulegen, dabei möglicherweise auftretende Risiken abzusehen und diesen effektiv vorzubeugen, beispielsweise durch Unterrichtung der Betreiber und Überwachung des Systembetriebs.[38] Insbesondere kann der Anbieter dafür sorgen, „dass das System **integrierten Betriebseinschränkungen** unterliegt, über die sich das System selbst **nicht hinwegsetzen** kann“,[39] damit im Rahmen des bestimmungsgemäßen Trainings keine Veränderung am System im Sinne des Art 25 KI-VO eintritt.[40]

Die besprochene Verschiebung der Rollen hätte nach Art 25 Abs 2 KI-VO zur Folge, dass der ursprüngliche Anbieter, der „**Erstanbieter**“, nicht mehr für die Erfüllung der anbieterspezifischen Pflichten der Verordnung einstehen muss. Für diesen Fall sind zwischen dem Anbieter und dem Betreiber vernünftigerweise nähere Vereinbarungen zu treffen: „Dieser Erstanbieter arbeitet eng mit neuen Anbietern zusammen, stellt die erforderlichen Informationen zur Verfügung und sorgt für den vernünftigerweise zu erwartenden technischen Zugang und sonstige Unterstützung, die für die Erfüllung der in dieser Verordnung festgelegten Pflichten, insbesondere in Bezug auf die Konformitätsbewertung von Hochrisiko-KI-Systemen, erforderlich sind.“

Aus praktischen Gründen – da es im gegenständlichen Projekt insbesondere um den Einsatz von KI-Systemen durch Strafverfolgungsbehörden geht – soll im Folgenden das System der verschiedenen Pflichten nach der KI-VO vorwiegend **aus der Sicht des Betreibers** am Beispiel von Polizeibehör-

36 „Entlang der KI-Wertschöpfungskette liefern häufig mehrere Parteien KI-Systeme, Instrumente und Dienstleistungen, aber auch Komponenten oder Prozesse, die vom Anbieter zu diversen Zwecken in das KI-System integriert werden; dazu gehören das **Trainieren, Neutrainieren, Testen und Bewerten von Modellen, die Integration in Software oder andere Aspekte der Modellentwicklung**. Diese Parteien haben eine wichtige Rolle in der Wertschöpfungskette gegenüber dem Anbieter des Hochrisiko-KI-Systems, in das ihre KI-Systeme, Instrumente, Dienste, Komponenten oder Verfahren integriert werden, und sollten in einer schriftlichen Vereinbarung die Informationen, die Fähigkeiten, den technischen Zugang und die sonstige Unterstützung nach dem allgemein anerkannten Stand der Technik bereitstellen, die erforderlich sind, damit der Anbieter die in dieser Verordnung festgelegten Pflichten vollständig erfüllen kann, ohne seine eigenen Rechte des geistigen Eigentums oder Geschäftsgeheimnisse zu gefährden.“

37 Siehe Art 13 Abs 3 lit c KI-VO.

38 Dabei stellen sich freilich auch datenschutzrechtliche Fragen, inwieweit der Anbieter durch den Betreiber ermittelte personenbezogene Daten einsehen und verarbeiten darf.

39 ErwGr 73 KI-VO.

40 Zur Problematik des „Trainierens“ weitaus genauer unten.

den dargestellt werden. Dabei werden die jeweiligen Themenbereiche im Einzelnen behandelt. Generell sind die Pflichten der Betreiber von KI-Systemen in Art 26 KI-VO geregelt; zudem ist der Betreiber Adressat der Vorschriften über die Grundrechte-Folgenabschätzung nach Art 27; daneben treffen den Betreiber bestimmte Pflichten im Zusammenhang mit der Meldung „schwerwiegender Vorfälle" bei der Marktüberwachungsbehörde gemäß Art 73, der sich jedoch primär auf den Anbieter bezieht;[41] schließlich ist es der Betreiber, dem gegenüber eine betroffene Person das Recht hat, im Sinne des Art 86 die „Erläuterung der Entscheidungsfindung im Einzelfall" zu verlangen.

Speziell zu Strafverfolgungsbehörden ist zu bemerken, dass sie von gewissen Verpflichtungen zur Offenlegung oder Übermittlung von Informationen ausgenommen sind, sofern es sich dabei um „**sensible operative Daten**" handelt.[42] Laut der Definition des Art 3 Z 38 KI-VO sind das „operative Daten im Zusammenhang mit Tätigkeiten zur Verhütung, Aufdeckung, Untersuchung oder Verfolgung von Straftaten, deren Offenlegung die **Integrität von Strafverfahren gefährden** könnte". Dies dient sowohl der Integrität der Strafverfolgung als auch den Grundrechten der betroffenen Personen, liegt also im Interesse der Rechtsstaatlichkeit. Der Begriff der „sensiblen operativen Daten" darf nicht verwechselt werden mit dem Begriff der „sensiblen Daten", als welche häufig „besondere Kategorien personenbezogener Daten" nach Art 9 der Datenschutz-Grundverordnung (DSGVO) bezeichnet werden, beispielsweise Daten über die Religionszugehörigkeit, über die Gesundheit oder die Sexualität eines Menschen sowie biometrische Daten.[43]

3.1. technische und organisatorische Maßnahmen

Art 26 Abs 1 KI-VO verpflichtet Betreiber von Hochrisiko-KI-Systemen dazu, „geeignete technische und organisatorische Maßnahmen" zu treffen, um eine Verwendung der KI gemäß den vom Anbieter bereitgestellten Betriebsanleitungen zu gewährleisten, um die **menschliche Aufsicht** über das KI-System zu garantieren und um für die Aufbewahrung der vom KI-System automatisch erzeugten **Protokolle** Sorge zu tragen.[44] Bei **Betriebsanleitungen** handelt es sich gemäß Art 3 Z 15

41 Dazu näher Art 26 Abs 5 KI-VO.

42 Siehe zum Beispiel Art 5 Abs 4, Art 26 Abs 5, Art 72 Abs 2 KI-VO.

43 Vgl. Artikel 10 der Richtlinie (EU) 2016/680 (genannt Datenschutzrichtlinie für Polizei und Justiz, DSRL-PJ); Art 3 Z 37 KI-VO.

44 Dazu ErwGr 91 KI-VO: „Angesichts des Charakters von KI-Systemen und der Risiken für die Sicherheit und die Grundrechte, die mit ihrer Verwendung verbunden sein können, ist es angezeigt, besondere Zuständigkeiten für die Betreiber festzulegen, auch im Hinblick darauf, dass eine angemessene Beobachtung der Leistung eines KI-Systems unter Realbedingungen sichergestellt werden muss. Die Betreiber sollten insbesondere geeignete technische und organisatorische Maßnahmen treffen, um sicherzustellen, dass sie Hochrisiko-KI-Systeme gemäß den Betriebsanleitungen verwenden, und es sollten bestimmte andere Pflichten in Bezug auf die Überwachung der Funktionsweise der KI-Systeme und gegebenenfalls auch Aufzeichnungspflichten festgelegt werden. Darüber hinaus sollten die Betreiber sicherstellen, dass die Personen, denen die Umsetzung der Betriebsanleitungen und die menschliche Aufsicht

KI-VO um „die Informationen, die der Anbieter bereitstellt, um den Betreiber insbesondere über die Zweckbestimmung und die ordnungsgemäße Verwendung eines KI-Systems zu informieren"; sie werden in Art 13 näher geregelt und beziehen sich unter anderem auf die Zweckbestimmung und die Merkmale des KI-Systems, auf Aspekte der Cybersicherheit, auf mögliche Risiken für Grundrechte, auf die technischen Fähigkeiten des Systems, dessen Leistung in Bezug auf bestimmte Personen, die erforderlichen Eingabe- und die jeweiligen Trainingsdaten, die entsprechenden Methoden zur Interpretation von Ausgaben, Fragen der Protokollierung, die Gewährleistung menschlicher Aufsicht sowie die benötigten Ressourcen wie Hardware und Software samt Updates und anderen Wartungs- und Pflegemaßnahmen.

Die Pflicht, geeignete technische und organisatorische Maßnahmen zwecks Konformität mit der KI-VO zu treffen, bedeutet einerseits, die beteiligten MitarbeiterInnen angemessen zu schulen, und andererseits, spezielle Abteilungen, Kommunikationswege und Verfahren zu schaffen bzw. weiterzuentwickeln. Wie genau dies zu geschehen habe, wird von Art 26 KI-VO den betreffenden Organisationen freigestellt bzw. im – hier interessierenden – Falle staatlicher Stellen im Lichte des unionsrechtlichen Grundsatzes der **Verfahrensautonomie** dem nationalen Recht überlassen; die Rede ist in Abs 3 genauer von der „Freiheit der Betreiber bei der Organisation ihrer eigenen Ressourcen und Tätigkeiten zur Wahrnehmung der vom Anbieter angegebenen Maßnahmen der menschlichen Aufsicht". Unionsrechtlich ist allerdings jedenfalls gefordert, dass die von der Behörde bzw. vom nationalen Gesetzgeber ergriffenen organisatorischen, personellen und prozeduralen Maßnahmen die **effektive Um- und Durchsetzung der KI-VO** gewährleisten. Der unionsrechtliche Grundsatz der **Äquivalenz** verlangt ferner, dass „Verfahren, in denen Unionsrecht vollzogen wird, [...] nicht weniger günstig gestaltet sein [dürfen] als Verfahren, die nur innerstaatliches Recht betreffen".[45] Das bedeutet, dass bei den von der KI-VO vorgegebenen technischen und organisatorischen Maßnahmen nicht weniger sorgfältig vorgegangen werden darf als bei solchen Maßnahmen, die bereits nach nationalem Recht für das Funktionieren der Polizeibehörden getroffen werden, ob nun ein Bezug zu KI besteht oder nicht.[46]

gemäß dieser Verordnung übertragen wurde, über die erforderliche Kompetenz verfügen, insbesondere über ein angemessenes Niveau an KI-Kompetenz, Schulung und Befugnis, um diese Aufgaben ordnungsgemäß zu erfüllen. Diese Pflichten sollten sonstige Pflichten des Betreibers in Bezug auf Hochrisiko-KI-Systeme nach Unionsrecht oder nationalem Recht unberührt lassen."

45 *Budischowsky* in *Jaeger/Stöger* (Hrsg), EUV/AEUV Art 197 AEUV (Stand 1.7.2023, rdb.at), Rz 7; zu all dem ebd., Rz 5 ff; *Klamert/Stöger* in *Jaeger/Stöger* (Hrsg), EUV/AEUV Art 4 EUV (Stand 1.1.2023, rdb.at), Rz 108 f; *Nehl* in *Jaeger/Stöger* (Hrsg), EUV/AEUV Art 19 EUV (Stand 1.3.2020, rdb.at), Rz 32; *Klamert*, EU-Recht (2015), Rz 108 ff, 743 ff, mit weiteren Nachweisen.

46 Zu den bereits bestehenden bzw. gesetzlich vorgesehenen Bildungs- und Qualitätssicherungsmaßnahmen auf dem Gebiet des Verfassungsschutzes: *Figl* in *Heißl/Figl* (Hrsg), SNG², § 2, Rz 27 ff.

Auch in diesem Zusammenhang sind die Ethik-Leitlinien der Unabhängigen Hochrangigen Expertengruppe von Interesse. Dort finden sich Vorschläge zur ethisch verantwortungsvollen Einrichtung eines Unternehmens, welches ein KI-System betreibt, gegliedert nach Organisationsebenen und Aufgabenbereichen.[47] Zwar sollte eine Polizeibehörde als solche keinesfalls nach denselben Kriterien wie ein Unternehmen beurteilt werden, jedoch dürften die Ausführungen der Expertengruppe dennoch für bestimmte Aspekte der KI-Verwendung in diesem Bereich lehrreich sein. Beispielsweise kommt der „Rechtsabteilung" folgende Rolle zu:

> „Die Abteilung für Unternehmensverantwortung überwacht die Umsetzung der Bewertungsliste und deren Weiterentwicklung, die erforderlich ist, um die Beurteilung an technische oder regulatorische Änderungen anzupassen. Sie aktualisiert die Vorschriften oder internen Richtlinien für KI-Systeme und stellt sicher, dass die Nutzung der KI-System im Einklang mit den bestehenden Rechts- und Verwaltungsvorschriften sowie den Unternehmenswerten erfolgt."

Zudem wird unter anderem eine eigene Abteilung für die Überprüfung der Ergebnisse des KI-Systems und für die Korrektur von Fehlern, Mängeln oder Verzerrungen empfohlen:

> „Die Abteilung für Qualitätssicherung (oder eine gleichwertige) überprüft die Ergebnisse der Arbeit mit der Bewertungsliste und ergreift Maßnahmen um ein Problem weiter oben zu eskalieren, wenn ein Ergebnis nicht zufriedenstellend ist oder wenn unvorhergesehene Ergebnisse festgestellt worden sind."

Schließlich solle es auch eine spezielle Abteilung für die Beschaffung von KI-Systemen geben. Im staatlichen Bereich ist diese Aufgabe durch das unionsrechtlich geprägte **Vergaberecht** geregelt und auf eine eigene Institution, die BundesbeschaffungsGmbH, übertragen.[48] Inhaltlich ist hier insbesondere auf den „Beschaffungsleitfaden für sichere AI" zu verweisen.[49]

Eine prinzipielle Bedeutung kommt in der KI-VO der „KI-Kompetenz" im Sinne des Art 4 zu. Demnach sind alle Personen bzw. im Falle der Polizeibehörde alle Bediensteten, die Umgang mit KI-Systemen haben oder auf deren Arbeit KI-Systeme einen Einfluss haben, angemessen auf die

47 *Unabhängige Hochrangige Expertengruppe für künstliche Intelligenz*, Ethik-Leitlinien für eine vertrauenswürdige KI (2019), Tabelle zu Rz 116.

48 Bundesgesetz über die Errichtung einer Bundesbeschaffung Gesellschaft mit beschränkter Haftung (BB-GmbH-Gesetz), BGBl. I Nr. 39/2001.

49 Siehe *Kieseberg/Kaltenbrunner/Temper/Tjoa*, Beschaffung sicherer AI-Systeme – Herausforderungen & Praktischer Leitfaden (http://ffhoarep.fh-ooe.at/bitstream/123456789/1726/1/FFH2023_244.pdf, Abruf 17.11.2024).

verschiedenen möglichen KI-relevanten Situationen vorzubereiten, um das Risiko, welches von einem unverständigen Gebrauch solcher Systeme ausgeht, möglichst weitgehend zu reduzieren:

> „Die Anbieter und Betreiber von KI-Systemen ergreifen Maßnahmen, um nach besten Kräften sicherzustellen, dass ihr Personal und andere Personen, die in ihrem Auftrag mit dem Betrieb und der Nutzung von KI-Systemen befasst sind, über ein ausreichendes Maß an **KI-Kompetenz** verfügen, wobei ihre technischen Kenntnisse, ihre Erfahrung, ihre Ausbildung und Schulung und der Kontext, in dem die KI-Systeme eingesetzt werden sollen, sowie die Personen oder Personengruppen, bei denen die KI-Systeme eingesetzt werden sollen, zu berücksichtigen sind."

Der Begriff der KI-Kompetenz im Sinne dieses Grundsatzes wird wiederum in Art 3 Z 56 KI-VO definiert als „die Fähigkeiten, die Kenntnisse und das Verständnis, die es Anbietern, Betreibern und Betroffenen unter Berücksichtigung ihrer jeweiligen Rechte und Pflichten im Rahmen dieser Verordnung ermöglichen, KI-Systeme sachkundig einzusetzen sowie sich der **Chancen und Risiken von KI und möglicher Schäden**, die sie verursachen kann, **bewusst** zu werden."[50]

Das bedeutet vor allem, dass unbedingt **Schulungs- und Weiterbildungsmaßnahmen** für die betreffenden MitarbeiterInnen ergriffen werden müssen, was freilich auch in budgetärer Hinsicht bedacht werden muss. Der Begriff „KI-Kompetenz" bezieht sich einerseits auf die **technischen Details**, deren Kenntnis für die behördliche Funktion notwendig ist – beispielsweise die richtige Art der Bedienung, Vorkehrungen der Cybersicherheit, grundlegende Informationen über die Arbeit mit Daten und Wahrscheinlichkeiten etc. –, andererseits jedoch auch gleichermaßen auf **Werte und ethische Aspekte bzw. Risiken** des Einsatzes von KI in einem so hochsensiblen Bereich wie der Strafverfolgung. Bildungsprogramme für Bedienstete sollen daher insbesondere Fragen wie etwa folgende gründlich behandeln: Wie kann der Einsatz von KI mit der Achtung von Grundrechten aller betroffenen Personen bzw. BürgerInnen vereinbar werden? Wie schafft man es, einerseits allgemein menschliche, andererseits durch den KI-Einsatz möglicherweise zusätzlich erzeugte **Vorurteile** auszuräumen? Wie soll in einer Behörde die Zusammenarbeit von technischem Personal, insbe-

50 Dazu näher ErwGr 20 KI-VO: „Um den größtmöglichen Nutzen aus KI-Systemen zu ziehen und gleichzeitig die Grundrechte, Gesundheit und Sicherheit zu wahren und eine demokratische Kontrolle zu ermöglichen, sollte die KI-Kompetenz Anbieter, Betreiber und betroffene Personen mit den notwendigen Konzepten ausstatten, um fundierte Entscheidungen über KI-Systeme zu treffen. Diese Konzepte können in Bezug auf den jeweiligen Kontext unterschiedlich sein und das Verstehen der korrekten Anwendung technischer Elemente in der Entwicklungsphase des KI-Systems, der bei seiner Verwendung anzuwendenden Maßnahmen und der geeigneten Auslegung der Ausgaben des KI-Systems umfassen sowie — im Falle betroffener Personen — das nötige Wissen, um zu verstehen, wie sich mithilfe von KI getroffene Entscheidungen auf sie auswirken werden. Im Zusammenhang mit der Anwendung dieser Verordnung sollte die KI-Kompetenz allen einschlägigen Akteuren der KI-Wertschöpfungskette die Kenntnisse vermitteln, die erforderlich sind, um die angemessene Einhaltung und die ordnungsgemäße Durchsetzung der Verordnung sicherzustellen."

sondere KI-SpezialistInnen, JuristInnen und Exekutivbediensteten funktionieren, wem fällt welcher Aufgabenbereich zu, wie können diese verschiedenen AkteurInnen einander durch ihre besonderen Kompetenzen ergänzen oder wechselseitig überprüfen – welche Kommunikationswege können sinnvoll vorgesehen werden, wie lassen sich Schwierigkeiten der „Übersetzung" zwischen juristisch, informatisch und polizeilich geschulten Personen reduzieren?

Passend dazu legt bereits beispielsweise § 11 Abs 1a SNG fest, dass Ermittlungen im Aufgabenbereich **Nachrichtendienst** durch **geeignete und besonders geschulte Bedienstete** wahrgenommen werden können.[51] Im Lichte des unionsrechtlichen Effektivitätsprinzips sollte diese besondere Schulung der Bediensteten dahingehend verstanden werden, dass sie auch sowohl die technischen wie die ethischen Aspekte der KI-Kompetenz im Sinne des Art 4 KI-VO umfasst, insofern diese Bediensteten in irgendeiner Weise mit KI-Systemen arbeiten. Erwägenswert wäre auch die Schaffung von Stellen für Personen, welche speziell für die Aufarbeitung und Klärung ethischer Probleme zuständig sind, zum Beispiel zur Minderung der möglichen Risiken der Diskriminierung, der Intransparenz, fehlenden oder unzureichenden rechtlichen Gehörs (= Anhörung der Betroffenen), …

3.2. menschliche Aufsicht

Art 14 Abs 4 KI-VO zum Thema „menschliche Aufsicht" über KI-Systeme in hochriskanten Bereichen kann als eine *lex specialis* zu Art 4 über „KI-Kompetenz" angesehen werden: Diese Vorschrift bestimmt nämlich, dass Personen, welche im Sinne der Verordnung mit der Aufsicht über KI-Systeme betraut sind, in der Lage sein sollen:

> a) die einschlägigen Fähigkeiten und Grenzen des Hochrisiko-KI-Systems angemessen zu **verstehen** und seinen Betrieb ordnungsgemäß zu überwachen, einschließlich in Bezug auf das Erkennen und Beheben von **Anomalien, Fehlfunktionen und unerwarteter Leistung**;
> b) sich einer möglichen Neigung zu einem automatischen oder übermäßigen Vertrauen in die von einem Hochrisiko-KI-System hervorgebrachte Ausgabe („**Automatisierungsbias**") bewusst zu bleiben, insbesondere wenn Hochrisiko-KI-Systeme Informationen oder Empfehlungen ausgeben, auf deren Grundlage natürliche Personen Entscheidungen treffen;
> c) die Ausgabe des Hochrisiko-KI-Systems richtig zu **interpretieren**, wobei beispielsweise die vorhandenen Interpretationsinstrumente und -methoden zu berücksichtigen sind;

51 Vgl. *Heißl* in *Heißl/Figl* (Hrsg), SNG², § 11, Rz 151.

d) in einer bestimmten Situation zu beschließen, das Hochrisiko-KI-System **nicht zu verwenden** oder die Ausgabe des Hochrisiko-KI-Systems **außer Acht zu lassen, außer Kraft zu setzen oder rückgängig zu machen**;

e) in den Betrieb des Hochrisiko-KI-Systems einzugreifen oder den Systembetrieb mit einer „**Stopptaste**“ oder einem ähnlichen Verfahren zu unterbrechen, was dem System ermöglicht, in einem sicheren Zustand zum Stillstand zu kommen.“

Dazu normiert Art 26 Abs 2 KI-VO: „Die Betreiber übertragen natürlichen Personen, die über die erforderliche Kompetenz, Ausbildung und Befugnis verfügen, die menschliche Aufsicht und lassen ihnen die erforderliche Unterstützung zukommen.“ Somit zählt die Gewährleistung menschlicher Aufsicht zu den zentralen Aufgaben des Betreibers nach der Verordnung – es handelt sich um eine Verpflichtung, welche nicht nur punktuell oder unter bestimmten Bedingungen, sondern laufend und während des gesamten Betriebs des KI-Systems erfüllt werden muss. Dabei sind die nötigen Kompetenzen der Aufsichtsperson in Art 14 Abs 4 teils technischer Natur – beispielsweise die Betätigung einer „Stopptaste“ oder die Einschätzung, ob eine „Anomalie“ oder ein normaler Betrieb vorliegt, oder die Interpretation von Ausgaben des KI-Systems, verstanden als Übersetzung mathematischer Werte wie zB Wahrscheinlichkeiten in natürliche Sprache –, teils handelt es sich jedoch auch um **Reflexionsleistungen** des betreffenden Organs, wobei sich typischerweise Fragen wie folgende stellen dürften: Führt die Verwendung des KI-Systems dazu, dass bestimmte Ausgaben unkritisch hingenommen werden? Inwieweit ist das Vertrauen in die Zuverlässigkeit der Ergebnisse des Systems gerechtfertigt („Automatisierungsbias“)? Welcher Teil der Überlegungen, die für Ermittlungen relevant sind, kann auf die Technologie „ausgelagert“ werden, welche Entscheidungen oder Gedankengänge müssen ihrer Natur nach bzw. im Lichte des aktuellen Standes der Technik bei menschlichen Organen verbleiben? Welche Situationen sind besonders anfällig für Vorurteile, wie lassen sich diese erkennen und beheben? Gibt es womöglich Fälle, in denen KI-Systeme besser überhaupt nicht verwendet werden sollen? Können etwaige Fehler, die im KI-System bzw. durch dieses auftreten, rückgängig gemacht werden? Von welcher Tragweite können bestimmte mögliche Verzerrungen sein?

Darüber hinaus stellt sich die grundlegende Frage, ob **Rechtskonformität** von KI-Systemen überhaupt ohne menschliche Aufsicht möglich sei, also selbst dann, wenn das System einwandfrei funktioniert und in technischer Hinsicht derart ausgereift ist, dass es einerseits die eigenen Fehler selbst ausbessert und insoweit keiner menschlichen Intervention bedarf, andererseits ohne jegliche Vorkenntnisse für das menschliche Organ hinreichend bedienbar und interpretierbar ist. So gibt es bei-

spielsweise im Datenschutzrecht die Konzepte „privacy by default“ und „privacy by design“.[52] Sie legen es nahe, dass Datenverarbeitungssysteme bereits aufgrund ihrer technischen Gestaltung automatisch rechtskonform sein könnten. Das Problem ist dabei, dass die „Einprogrammierung“ von Rechtsnormen prinzipielle Schwierigkeiten aufwirft: Zwar können durch die informationstechnische Umsetzung rechtlicher Vorschriften bedeutende Fortschritte erreicht werden, jedoch beinhalten zahlreiche Rechtsnormen als solche sprachliche Unklarheiten oder Offenheit; teils in semantischer Hinsicht wie im Falle von vagen Formulierungen oder Generalklauseln, teils in syntaktischer Hinsicht, also bereits auf der Ebene der logischen Struktur des Satzes, beispielsweise aufgrund eines mehrdeutigen Gebrauchs des Wortes „oder“.[53] Zu denken ist dabei auch an Formulierungen, welche politischen Kompromissen geschuldet sind und daher in sprachlicher Hinsicht nicht zufriedenstellend sind.[54] In all diesen Fällen muss man, um die Norm einem KI-System „einzuprogrammieren“, zuerst durch menschliche oder KI-basierte **Interpretation** aus der unpräzisen Norm einen hinreichend präzisen Code erstellen. Dabei wird aber womöglich die ursprüngliche Bedeutung eingeengt bzw. modifiziert: Das formalisierte Resultat könnte nicht mit dem Willen des Gesetzgebers übereinstimmen, dieser deckt sich nicht notwendig mit der durch Interpretation erzeugten, für Computersysteme brauchbaren Regel.[55] Deshalb ist die Anwendung oder Auslegung von Rechtsnormen nach wie vor eine Aufgabe, für welche menschliches Denken essenziell ist und die somit nicht restlos auf Computer übertragbar ist. In diesem Sinne lässt sich die Aufgabe der Herstellung und Überprüfung der Rechtskonformität von KI-Systemen als **menschliche Aufsicht im weiteren Sinne** bezeichnen.

Hinsichtlich der näheren Modalitäten für die menschliche Aufsicht kann wiederum auf die Ethikleitlinien der Hochrangigen Expertengruppe verwiesen werden, welche verschiedene Modelle der „menschlichen Aufsicht“ je nach der Rolle des beaufsichtigenden menschlichen Akteurs anführen:

> „Die menschliche Aufsicht hilft, dafür zu sorgen, dass ein KI-System die menschliche Autonomie nicht untergräbt oder sich sonst nachteilig auswirkt. Die Aufsicht kann durch Lenkungs- und Kontrollmechanismen wie die Gewährleistung der interaktiven Einbindung eines Menschen („*Human-in-the-Loop*“), der Überprüfung und Kontrolle durch einen Menschen („*Human-on-the-Loop*“) oder der Gesamtsteuerung durch einen Menschen („*Human-in-Command*“) erreicht werden. „Human-in-the-Loop“ (HITL) bezieht sich auf die Fähigkeit des Menschen, in jeden Entscheidungszyklus des Systems einzugreifen, was in vie-

52 Vgl. ErwGr 78 DSGVO und Art 25 DSGVO.
53 Vgl. *Ashley*, Artificial Intelligence and Legal Analytics, 39 ff.
54 Vgl. ebd., 40.
55 „The problem is that the version of a statutory rule selected for formalization into a logic programming language like Prolog is not necessarily the one that the legislature intended.“ (ebd., 45)

len Fällen weder möglich noch wünschenswert wäre. „Human-on-the-Loop“ (HOTL) bedeutet die Möglichkeit des Menschen, in den Entwurfszyklus des Systems einzugreifen und den Systembetrieb zu überwachen. „Human-in-Command“ (HIC) bedeutet die Möglichkeit, den Gesamtbetrieb des KI-Systems (einschließlich seiner weiteren wirtschaftlichen, gesellschaftlichen, rechtlichen und ethischen Auswirkungen) zu beaufsichtigen, sowie die Fähigkeit zu entscheiden, wann und wie das System in einer bestimmten Situation eingesetzt werden soll. Dies beinhaltet auch die Entscheidungsmöglichkeit, ein KI-System in einer bestimmten Situation nicht einzusetzen, beim Einsatz des Systems ein bestimmtes Maß an menschlichem Ermessen zuzulassen oder sicherzustellen, dass eine vom System getroffene Entscheidung außer Kraft gesetzt werden kann. Des Weiteren muss sichergestellt werden, dass die Durchsetzungsbehörden stets in der Lage sind, die Aufsicht im Einklang mit ihrem jeweiligen Auftrag auszuüben. Aufsichtsmechanismen sind möglicherweise in unterschiedlichen Graden notwendig, um andere Sicherheits- und Kontrollmaßnahmen je nach Anwendungsbereich und potenziellem Risiko des KI-Systems zu unterstützen. Für ein ansonsten gleiches System muss gelten: Je weniger Aufsicht ein Mensch über ein KI-System ausüben kann, desto ausführlicher muss es zuvor getestet werden und desto strenger muss die Lenkung und Kontrolle sein.“[56]

Die in diesen Ausführungen vorgenommenen Unterscheidungen legen es nahe, dass die Kriterien für die nähere Umsetzung des Konzepts „menschliche Aufsicht“ zu einem großen Teil **technischer Natur** sind: Es ist schließlich eine Frage der Spezifikationen des KI-Systems, welche Eingriffsmöglichkeiten menschlichen Personen zur Verfügung stehen und wie weitreichend sie sind. Eine bloße Beschreibung der jeweiligen Handlungsoptionen der beaufsichtigenden Person kann der rechtlichen Beurteilung, ob diese ausreichend seien, nicht vorgreifen. Die „interaktive Einbindung eines Menschen“ (*human in the loop*) mit der Möglichkeit, „in jeden Entscheidungszyklus des Systems einzugreifen“, ist zwar im Sinne der Leitlinien „in vielen Fällen weder möglich noch wünschenswert“; doch ist dadurch nicht ausgeschlossen, dass es sehr wohl bestimmte Fälle geben könnte, in denen ein solches besonders granulares Konzept in rechtlicher Hinsicht angezeigt ist. Insbesondere könnten in sensiblen Bereichen gewisse Situationen eintreten, bei denen im Zweifel zum Schutz der Grundrechte das jeweils feinkörnigere Modell für die menschliche Aufsicht gewählt werden sollte. **Rechtlich** betrachtet geht es bei der „menschlichen Aufsicht“ darum, möglichen Fehlern, Fehlschlüssen und Vorurteilen im KI-System keinen entscheidenden Einfluss auf behördliches Handeln gegenüber betroffenen Personen zu erlauben – wann so etwas der Fall sein könnte, muss erstens auf-

56 *Unabhängige Hochrangige Expertengruppe für künstliche Intelligenz*, Ethik-Leitlinien für eine vertrauenswürdige KI, Rz 65.

grund der Besonderheiten des jeweiligen Verwendungskontextes (bzw. Verwendungszwecks) und zweitens im Rahmen des **Einzelfalls** bei sorgfältiger Betrachtung sämtlicher bekannter Umstände beurteilt werden. Nicht zuletzt ist in Erinnerung zu behalten, dass in Fällen außergewöhnlicher, über die reguläre Arbeit hinausgehender Unsicherheit nicht nur die Möglichkeit, sondern auch die Pflicht bestehen kann, von der Verwendung eines KI-Systems überhaupt abzusehen und ausschließlich von den herkömmlichen, menschlichen Fähigkeiten Gebrauch zu machen.

Im Übrigen ist die Gestaltung, Herstellung oder Programmierung der KI-Systems dahingehend, dass die menschliche Aufsicht im Sinne der Verordnung gewährleistet ist, eine Aufgabe des **Anbieters**. Sie ist in Art 14 KI-VO geregelt[57] und bedeutet unter anderem, dass der Betreiber bzw. das von ihm eingesetzte Organ durch die technischen Merkmale der KI-Systems – zum Beispiel eine „**Mensch-Maschine-Schnittstelle**" im Sinne des Art 14 Abs 1 – und entsprechende Informationen in der **Betriebsanleitung** vom Anbieter in die Lage gesetzt wird, die im bereits zitierten Abs 4 angeführten Fähigkeiten zu erwerben.[58] Die zu implementierenden Aufsichtsmöglichkeiten müssen im Sinne des Abs 3 „den Risiken, dem Grad der Autonomie und dem Kontext der Nutzung des Hochrisiko-KI-Systems angemessen sein". Sie können einerseits bereits in das KI-System „eingebaut" sein, andererseits jedoch auch einer näheren Umsetzung durch den Betreiber unterliegen, sofern dieser über die nötigen technischen Kompetenzen verfügt. Beispielsweise könnte das beauftragte Technologieunternehmen als Anbieter der Polizeibehörde verschiedene Möglichkeiten oder Modelle der menschlichen Aufsicht zur Wahl stellen, welche je nach Verwendungskontext von besonders geschulten Organen zu treffen ist (beispielsweise in Form einer Wahlmöglichkeit zwischen „*human in/on the loop*" und „*human in command*").[59]

Letztlich muss der Anbieter immer auch mit größter Sorgfalt erwägen, wann im Rahmen der Verwendung des von ihm bereitgestellten Systems durch den Betreiber etwas anders läuft als von ihm selbst geplant und wie in solchen Fällen vorzugehen sei. Beispielsweise können spontan oder durch menschliches Versehen verschiedene Arten von Programmfehlern auftreten, auch kann eine neuartige Situation auftreten, bei welcher die reguläre Betrieb des Systems auf Schwierigkeiten stößt. In Art 14 Abs 2 KI-VO ist die Rede von der „**vernünftigerweise vorhersehbaren Fehlanwen-**

57 Vgl. dazu auch ErwGr 73 KI-VO.
58 Vgl. Art 9 Abs 5 lit c, Art 13 Abs 3 lit d KI-VO.
59 ErwGr 72 KI-VO: „Unter anderem sollten Betreiber besser in der Lage sein, das richtige System auszuwählen, das sie angesichts der für sie geltenden Pflichten verwenden wollen, über die beabsichtigten und ausgeschlossenen Verwendungszwecke informiert sein und das KI-System korrekt und angemessen verwenden. Um die Lesbarkeit und Zugänglichkeit der in der Betriebsanleitung enthaltenen Informationen zu verbessern, sollten diese gegebenenfalls anschauliche Beispiele enthalten, zum Beispiel zu den Beschränkungen sowie zu den beabsichtigten und ausgeschlossenen Verwendungen des KI-Systems."

dung“.[60] Diese wird in Art 3 Z 13 definiert als „die Verwendung eines KI-Systems in einer Weise, die nicht seiner Zweckbestimmung entspricht, die sich aber aus einem vernünftigerweise vorhersehbaren menschlichen Verhalten oder einer vernünftigerweise vorhersehbaren **Interaktion mit anderen Systemen**, auch anderen KI-Systemen, ergeben kann“. Der Anbieter hat solche Situationen trotz der notwendigen Unsicherheit über Zukünftiges im voraus in Rechnung zu stellen, auch der Betreiber muss in weiterer Folge mit dieser Problematik gut vertraut sein und richtig auf solche Ausnahmefälle reagieren können.[61] Die Pflicht des Anbieters, die möglichen Risiken bei der Verwendung des KI-Systems durch den Anbieter vorherzusehen und ihnen wirksam zu begegnen, ist Teil des **Risikomanagements** gemäß Art 9 KI-VO, das nach Art 17 Abs 1 lit g wiederum Teil des **Qualitätsmanagements** ist.[62] „Das Risikomanagementsystem versteht sich als ein kontinuierlicher iterativer Prozess, der während des gesamten Lebenszyklus eines Hochrisiko-KI-Systems geplant und durchgeführt wird und eine regelmäßige systematische Überprüfung und Aktualisierung erfordert.“[63] Art 9 Abs 5 KI-VO verpflichtet den Anbieter des KI-Systems ausdrücklich dazu, „die technischen Kenntnisse, die Erfahrungen und de[n] Bildungsstand, die vom Betreiber erwartet werden können, sowie de[n] voraussichtliche[n] Kontext, in dem das System eingesetzt werden soll“, gebührend zu berücksichtigen.

Für **biometrische Fernidentifizierungssysteme** (Verweis auf Anhang III Z 1 lit a) enthält Art 14 Abs 5 KI-VO eine Sondervorschrift: Vorkehrungen zur Gewährleistung der menschlichen Aufsicht sollen „so gestaltet sein, dass außerdem der Betreiber keine Maßnahmen oder Entscheidungen allein aufgrund des vom System hervorgebrachten Identifizierungsergebnisses trifft, solange diese Identifizierung nicht von mindestens **zwei natürlichen Personen**, die die notwendige Kompetenz, Ausbildung und Befugnis besitzen, getrennt überprüft und bestätigt wurde.“ (**Vier-Augen-Prinzip**) Das „biometrische Fernidentifizierungssystem“ selbst ist gemäß Art 3 Z 41 KI-VO „ein KI-System, das dem Zweck dient, natürliche Personen ohne ihre aktive Einbeziehung und in der Regel aus der Ferne durch **Abgleich** der biometrischen Daten einer Person mit den in einer **Referenzdatenbank** gespeicherten biometrischen Daten zu identifizieren“. Dies umfasst sowohl Identifizierungen in Echtzeit als auch solche, die erst nachträglich durchgeführt werden.[64] **Biometrische Daten** (Art 3 Z 34 KI-VO) „mit speziellen technischen Verfahren gewonnene personenbezogene Daten zu den **physischen, physiologischen oder verhaltenstypischen Merkmalen** einer natürlichen Person, wie etwa

60 Vgl. Art 9 Abs 2 lit b KI-VO im Kontext des Risikomanagements des Anbieters.
61 Zur Informationspflicht des Anbieters in den Betriebsanleitungen Art 13 Abs 3 lit b sublit iii KI-VO.
62 Beides ist auf den Anbieter, nicht auf den Betreiber bezogen, vgl. Art 16 lit a und c KI-VO.
63 Art 9 Abs 2 KI-VO.
64 Für weitere einschlägige Definitionen vgl. Art 3 Z 35, 42, 43 KI-VO.

Gesichtsbilder oder daktyloskopische Daten".[65] Darunter fallen also beispielsweise solche Daten, die durch die Videoüberwachung oder die Abnahme von Fingerabdrücken gesammelt werden;[66] relevant sind aber auch privat genutzte Anwendungen (Apps), in denen Nutzer Gesundheitsdaten, Informationen über ihr Essverhalten, ihr Gewicht oder ihre körperlichen Bedürfnisse preisgeben, gleichermaßen Uhren, welche den Puls beim Laufen messen; ein Profilbild in einem sozialen Netzwerk kann ebenfalls ein biometrisches Datum sein, sofern es das Gesicht oder die Figur der Person realistisch darstellt; Informationen über Verletzungen wie Narben oder eine gebrochene Nase bei einem Verdächtigen oder möglichen Opfer sind biometrische Daten; ferner ließe sich diskutieren, ob Daten über das Sexualleben oder Liebesleben eines Menschen – welche im Rahmen einer „social network analysis" oder eines Mitlesens von Chat-Nachrichten durchaus anfallen können – aufgrund ihres untrennbaren Zusammenhangs mit körperlichen Zuständen, Bedürfnissen, Gewohnheiten unter die Kategorie der „biometrischen Daten" fallen. Wird in diesem Sinne die Identität aufgrund Daten der besprochenen Kategorie ermittelt, müssen also zwei hinreichend geschulte Personen diese Identifizierung überprüfen. Nun normiert der zweite Unterabsatz in Art 14 Abs 5 KI-VO wiederum eine Ausnahme: Das Erfordernis der Überprüfung durch zwei Personen gilt unter anderem nicht im Bereich der **Strafverfolgung**, „wenn die Anwendung dieser Anforderung nach Unionsrecht oder nationalem Recht unverhältnismäßig wäre."

Nach der hier vertretenen Ansicht ist ein solches Erfordernis **im Rahmen der StPO, des SPG und des SNG nicht unverhältnismäßig**. Zum einen ist bereits nach dem österreichischen Strafprozessrecht für bestimmte Ermittlungsmaßnahmen, welche auf besonders intensive Weise Datenverarbeitungen – beispielsweise den automationsunterstützten Datenabgleich („Rasterfahndung") nach § 141 StPO – oder verdeckte Eingriffe in das Kommunikationsgeheimnis bzw. in die Privatsphäre beinhalten,[67] die Überwachung bzw. Überprüfung der behördlichen Tätigkeit durch den **Rechtsschutzbeauftragten** (RSB) nach § 147 StPO vorgesehen. Es liegt in seiner Funktion, „die gesamten Ergebnisse" der Ermittlungsmaßnahme „einzusehen und anzuhören, bevor diese zum Akt genommen werden. Er ist ferner berechtigt, die Vernichtung von Ergebnissen oder Teilen von ihnen (§ 139 Abs. 4) zu beantragen und sich von der ordnungsgemäßen Vernichtung dieser Ergebnisse zu überzeugen. Das Gleiche gilt für die ordnungsgemäße Löschung von Daten, die in einen Datenabgleich einbezogen oder durch ihn gewonnen wurden. Beabsichtigt die Staatsanwaltschaft, einem solchen Antrag des Rechtsschutzbeauftragten nicht nachzukommen, so hat sie unverzüglich die Entschei-

65 ErwGr 15 KI-VO spricht von der „Erkennung physischer, physiologischer und verhaltensbezogener menschlicher Merkmale wie Gesicht, Augenbewegungen, Körperform, Stimme, Prosodie, Gang, Haltung, Herzfrequenz, Blutdruck, Geruch, charakteristischer Tastenanschlag".

66 Vgl. die Bestimmungen zum „Erkennungsdienst" nach §§ 64 ff SPG.

67 Für eine kritische Auseinandersetzung vgl. *Zerbes*, Spitzeln, Spähen, Spionieren, 159 ff.

dung des Gerichts einzuholen." (§ 147 Abs 4 StPO) Daher erscheint es durchaus konsequent, die Kompetenzen des RSB um die Aufgabe der menschlichen Aufsicht über KI-Systeme im Sinne des Art 14 KI-VO zu ergänzen bzw. eine dem RSB koordinierte Stelle dafür zu schaffen. Ähnliches gilt für den RSB nach §§ 91a ff SPG bzw. § 14 ff SNG, welchem zudem gemäß § 91d Abs 3 SPG bzw. § 16 Abs 1 SNG die allgemeine Aufsicht über geheime Verarbeitungen personenbezogener Daten zukommt. Zu bedenken ist außerdem, die Einbeziehung der **Datenschutzbehörde** (DSB), welche schon *de lege lata* nach § 142 Abs 3 und 4 StPO den automationsunterstützten Datenabgleich kontrolliert, in die menschliche Aufsicht zu normieren.[68] Damit wären die „zwei natürlichen Personen" im Sinne des Art 14 Abs 5 KI-VO ein Organ des RSB und eines der DSB.[69] Erst recht nicht unverhältnismäßig im Sinne dieser Vorschrift ist das Erfordernis der Überprüfung durch mindestens zwei Personen dann, wenn es sich dabei ausschließlich um **behördeninterne** Personen handelt.[70] – Ganz im Gegenteil, es könnte sogar auf ein Rechtsschutzdefizit hinauslaufen.[71]

3.3. datenbezogene Verantwortung des Betreibers

Gemäß Art 26 Abs 4 KI-VO „sorgen die Betreiber dafür, dass die Eingabedaten der Zweckbestimmung des Hochrisiko-KI-Systems entsprechen und ausreichend **repräsentativ** sind." Diese Pflicht trifft sie jedoch nur, „soweit die Eingabedaten ihrer Kontrolle unterliegen". Art 3 Z 33 KI-VO definiert **Eingabedaten** als „die in ein KI-System eingespeisten oder von diesem direkt erfassten Daten, auf deren Grundlage das System eine Ausgabe hervorbringt". Diese Art von Daten liegt, wie diese Textstellen nahelegen, grundsätzlich in der Sphäre des Betreibers: Er ist derjenige, der sie eingibt und in diesem Sinne verarbeitet. Eingabedaten stehen sozusagen auf der untersten Stufe der „KI-Wertschöpfungskette", weil sie unmittelbar zur Bedienung oder Nutzung des KI-Systems in einer konkreten Situation gehören („Anwenderebene"). Daneben kennt die KI-VO auch andere Kategorien von Daten, nämlich Trainings-, Validierungs- und Testdaten. Diese Kategorien von Daten sind den jeweiligen Rollen in der KI-VO zuzuordnen, es ist eine Abgrenzung der Zuständigkeiten vorzunehmen. So ist zunächst festzustellen, dass die Arbeit mit Eingabedaten vorwiegend in den Verantwortungsbereich der Betreiber fällt, während **Testdaten** eher nur für Anbieter relevant sind, weil sie gemäß Art 3 Z 32 KI-VO dazu „verwendet werden, um die erwartete Leistung dieses Systems vor dessen Inverkehrbringen oder Inbetriebnahme zu bestätigen", ihre Verarbeitung also der Nutzung

68 Siehe § 22 DSG zu den generellen Befugnissen der DSB.

69 Zu dem Verständnis der Gewaltenteilung im Kontext der Funktion des RSB: *Zerbes*, Spitzeln, Spähen, Spionieren, 158 f, mit weiteren Nachweisen.

70 Darunter zu verstehen sind auch **Datenschutzbeauftragte** im Sinne der Art 32 ff DSRL-PJ und § 57 DSG.

71 „Die Gefahr, dass die menschliche Aufsicht in einen reinen Formalismus abgleitet, ist weiterhin real und nicht zu unterschätzen." (*Martini* in *Martini/Wendehorst*, KI-VO, Art 14, Rz 125. Dort mit Hervorhebungen)

durch einen Betreiber vorgelagert ist. Die Frage nach dem systematischen Ort von Trainings- und Validierungsdaten lässt sich dagegen nicht pauschal beantworten. **Trainingsdaten** sind nach Art 3 Z 29 „Daten, die zum Trainieren eines KI-Systems verwendet werden, wobei dessen lernbare Parameter angepasst werden", **Validierungsdaten** nach Z 30 „Daten, die zur Evaluation des trainierten KI-Systems und zur Einstellung seiner nicht erlernbaren Parameter und seines Lernprozesses verwendet werden, um unter anderem eine Unter- oder Überanpassung zu vermeiden". Das bedeutet, dass die Verarbeitung von Daten dieser beiden Kategorien einerseits beim Anbieter bereits in der Entwicklungsphase des KI-Systems, also vor der Nutzung durch einen Betreiber stattfinden kann, andererseits aber auch während des Betriebs bzw. nach dem Inverkehrbringen auf dem Markt. In letzterem Fall ist wiederum danach zu fragen, ob und inwieweit das Trainieren oder Validieren (Bewerten, Einstellen) des KI-Systems eine Aufgabe des Anbieters, etwa mittels Update, oder des Betreibers ist. Hier sind verschiedene Konstellationen möglich: Der Anbieter könnte im einen Extremfall lediglich ein „Endnutzer" sein, welchem außer der normalen operativen Nutzung des KI-Systems in einem eng definierten Verarbeitungskontext keine weitere Aufgabe zukommt; im anderen Extremfall könnte sich der Anbieter seiner Zuständigkeit zum Trainieren, Warten etc. weitreichend begeben und sie zu möglichst großen Teilen dem Betreiber überlassen. Letzteres funktioniert selbstverständlich nur dann, wenn beim Betreiber eine besonders hohe technologische Kompetenz gewährleistet ist.[72] Jedenfalls ist in der **Betriebsanleitung** und den weiteren Unterlagen bzw. Vereinbarungen im Sinne des Art 13 KI-VO festzulegen, wer in welchem Abschnitt im Lebenszyklus des KI-Systems welche Aufgaben übernimmt und mit welchen Daten dabei gearbeitet wird.[73] Die Beantwortung der Frage: „**Wer trainiert das KI-System?**", richtet sich daher nach der konkreten Ausgestaltung des Verhältnisses zwischen Anbieter und Betreiber. Realistisch erscheint zum Beispiel, dass das „grundlegende" – also für den Betrieb überhaupt notwendige – Training des Systems dem Anbieter obliegt und dabei dem Betreiber die Möglichkeit oder auch die Verpflichtung anheimgestellt wird, das System im Zuge des Betriebs mit neuen Trainingsdaten weiter zu trainieren, also zu verfeinern, anzupassen, zu vervollständigen, zu korrigieren.[74]

So erscheint es rechtlich betrachtet durchaus notwendig, im Verlauf des Systembetriebs auftretenden Fehlern, Vorurteilen, Verzerrungen usw. durch korrigierendes Training im Einzelfall zu begeg-

72 Vgl. Art 9 Abs 5 KI-VO.

73 Laut Art 13 Abs 3 lit b sublit vi KI-VO enthält die Betriebsanleitung Informationen über die „Spezifikationen für die Eingabedaten oder sonstiger relevanter Informationen über die verwendeten Trainings-, Validierungs- und Testdatensätze, unter Berücksichtigung der Zweckbestimmung des Hochrisiko-KI-Systems".

74 Im Allgemeinen erfasst das **Qualitätsmanagement** des Anbieters gemäß Art 17 Abs 1 lit f KI-VO immer „Systeme und Verfahren für das Datenmanagement, einschließlich Datengewinnung, Datenerhebung, Datenanalyse, Datenkennzeichnung, Datenspeicherung, Datenfilterung, Datenauswertung, Datenaggregation, Vorratsdatenspeicherung und sonstiger Vorgänge in Bezug auf die Daten, die im Vorfeld und für die Zwecke des Inverkehrbringens oder der Inbetriebnahme von Hochrisiko-KI-Systemen durchgeführt werden".

nen. – Dies liegt insbesondere beim Einsatz durch eine staatliche Behörde im Interesse, den Grundsatz der **Datenrichtigkeit** (und -aktualität) gemäß § 37 Abs 1 Z 4 DSG zu erfüllen, welcher laut § 45 DSG ein durchsetzbares Recht der betroffenen Person begründet. Generell ist bei der Arbeit mit KI-Systemen stets nicht nur an die Vorgaben der KI-VO zu denken, sondern auch an diejenigen des Datenschutzrechts.[75] Ferner gefährden falsche oder einseitige Trainingsdaten die technische Funktionstüchtigkeit bzw. den normalen operativen Betrieb des KI-Systems und damit die Ermittlungen durch die Strafverfolgungsbehörde als solche, ganz abgesehen von der praktisch bedeutenden Möglichkeit einer rechtswidrigen Diskriminierung bestimmter Personen oder Personengruppen.[76]

Die zentrale Vorschrift über den richtigen Umgang mit Daten bei KI-Systemen im Allgemeinen ist Art 10 KI-VO mit dem Titel **„Daten und Daten-Governance".** Sie richtet sich an den **Anbieter** des KI-Systems[77] und thematisiert Trainings-, Validierungs- und Testdatensätze, nicht aber Eingabedaten: „Die Trainings-, Validierungs- und Testdatensätze müssen im Hinblick auf die Zweckbestimmung **relevant**, hinreichend **repräsentativ** und so weit wie möglich **fehlerfrei** und **vollständig** sein. Sie müssen die geeigneten statistischen Merkmale, gegebenenfalls auch bezüglich der Personen oder Personengruppen, für die das Hochrisiko-KI-System bestimmungsgemäß verwendet werden soll, haben. Diese Merkmale der Datensätze können auf der Ebene einzelner Datensätze oder auf der Ebene einer Kombination davon erfüllt werden." Auf technische bzw. statistische Einzelheiten kann hier nicht näher eingegangen werden; in diesem Zusammenhang sind jedenfalls die vielfältigen Punkte in Art 10 Abs 2 KI-VO zu beachten.

Geklärt werden muss zuerst, inwieweit ein **Betreiber**, beispielsweise eine Polizeibehörde, das fragliche KI-System **selbst „trainiert"**. Zwar ist auf den ersten Blick der Umgang mit „Eingabedaten" nach Art 26 Abs 4 eindeutig dem Betreiber und die Arbeit mit „Trainingsdaten" u.a. nach Art 10 (in Verbindung mit Art 16 lit a) dem Anbieter zugewiesen, doch scheint hier darüber hinaus eine **differenzierte** Herangehensweise angezeigt. Zunächst erkennt die Verordnung in Art 15 Abs 4 ausdrücklich solche Situationen und Sachverhalte an, in welchen das KI-System nach der Inbetriebnahme und während der Verwendung durch den Betreiber weiter trainiert wird: „Hochrisiko-KI-Systeme,

75 Vgl. ErwGr 69 KI-VO; *Schürmann*, KI und Recht: Die Datenschutz-Folgenabschätzung und die Entwicklung von KI-Systemen, DSB 2021, 320; *Rothmund-Burgwall*, Einsatz von KI als neuartige Datenverarbeitung?, Dako 2024, 55; *Helminger*, Datenschutzrechtliche Herausforderungen bei der Verwendung von Trainingsdaten, EALR 2022, 46; *Schemmel*, Grundrechte- und Datenschutz-Folgenabschätzung – zwei Seiten einer Medaille des Daten-Risikomanagements, CB 2024, 321.

76 Vgl. *Rachinger/Voithofer/Kettemann*, Jenseits der Daten-Monokulturen: Zur Notwendigkeit diversitätssensibler Trainingsdaten, jusIT 2024, 43; vgl. auch *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz (2019), 47 ff, 55 ff; allgemeiner *Holoubek*, Grundrechtsschutz vor neuen Herausforderungen, 21. ÖJT Band I/1, 76 ff.

77 Dieser ist gemäß Art 16 lit a KI-VO dafür verantwortlich, dass das KI-System den produktbezogenen Vorgaben der Verordnung entspricht.

die nach dem Inverkehrbringen oder der Inbetriebnahme weiterhin **dazulernen**, sind so zu entwickeln, dass das Risiko möglicherweise **verzerrter** Ausgaben, die künftige Vorgänge beeinflussen („**Rückkopplungsschleifen**"), beseitigt oder so gering wie möglich gehalten wird und sichergestellt wird, dass auf solche Rückkopplungsschleifen angemessen mit geeigneten Risikominderungsmaßnahmen eingegangen wird."[78] Ein Übergang der Verantwortlichkeiten im Sinne des Art 25 Abs 1 und 2 KI-VO, durch welchen der Betreiber zum Anbieter wird und der Erstanbieter von seinen Verpflichtungen frei wird, setzt eine „wesentliche Veränderung" des KI-Systems selbst oder eine Änderung von dessen Zweckbestimmung voraus. Nun kann aber der Anbieter im Rahmen seines Risiko- und Qualitätsmanagements vorsehen, dass das KI-System unter präzise definierten Bedingungen auf bestimmte Art und Weise in einem deutlich umgrenzten Umfang durch den Betreiber trainiert werden darf oder muss, und dies dem Betreiber in der Betriebsanleitung unter Angabe der vorgesehenen Modalitäten für dieses – insoweit eigenverantwortliche – Trainieren kommunizieren.[79] Ein solches Trainieren läge dann nach der hier vertretenen Ansicht gerade innerhalb der vom Anbieter festgelegten Konzeption und Zweckbestimmung des KI-Systems, mit der Konsequenz, dass die Bedingungen für den Rollentausch nach Art 25 KI-VO insofern nicht vorliegen. Art 26 Abs 4 und Art 10 KI-VO, welche den jeweiligen Rollen – Anbieter und Betreiber – typische Kategorien von Daten zuordnen, sind deshalb nicht dahingehend zu verstehen, dass sie im Falle des Trainierens durch einen Betreiber *per se* dessen Eigenschaft als Anbieter begründeten. Damit übereinstimmend wird auch in Art 43 Abs 4 zum Thema „Konformitätsbewertung" ausdrücklich bestimmt: „Bei Hochrisiko-KI-Systemen, die nach dem Inverkehrbringen oder der Inbetriebnahme weiterhin dazulernen, gelten Änderungen des Hochrisiko-KI-Systems und seiner Leistung, die vom Anbieter zum Zeitpunkt der ursprünglichen Konformitätsbewertung vorab festgelegt wurden und in den Informationen der technischen Dokumentation […] enthalten sind, nicht als wesentliche Veränderung".[80] – Umso dringlicher ist es deshalb, dass im Verhältnis zwischen Betreiber und Anbieter hinreichend genaue und präzise Vorgaben bzw. Vereinbarungen über die Abgrenzung der Zuständigkeiten, den Spielraum des Betreiber und die gründliche Zusammenarbeit getroffen werden. (Zumindest aus rechtspolitischer Sicht ist es erwünscht, dass einerseits die Polizeibehörde als Betreiber das KI-System selbst trainieren kann und andererseits die Verantwortlichkeit des Anbieters nicht zur Gänze auf

78 Vgl. ErwGr 88 KI-VO zur möglichen Rollenverteilung für das Trainieren im Zuge der Entwicklung des KI-Systems.

79 Siehe Art 13 Abs 3 lit c KI-VO.

80 Dabei wird verwiesen auf Anhang IV Z 2 lit f KI-VO: Die technische Dokumentation umfasst „gegebenenfalls detaillierte Beschreibung der vorab bestimmten Änderungen an dem KI-System und seiner Leistung mit allen einschlägigen Informationen zu den technischen Lösungen, mit denen sichergestellt wird, dass das KI-System die einschlägigen in Kapitel III Abschnitt 2 festgelegten Anforderungen weiterhin dauerhaft erfüllt." Vgl. auch ErwGr 128: „Änderungen, die den Algorithmus und die Leistung von KI-Systemen betreffen, die nach dem Inverkehrbringen oder der Inbetriebnahme weiterhin dazulernen — d. h., sie passen automatisch an, wie die Funktionen ausgeführt werden —, sollten jedoch keine wesentliche Veränderung darstellen, sofern diese Änderungen vom Anbieter vorab festgelegt und zum Zeitpunkt der Konformitätsbewertung bewertet wurden."

die Behörde übergeht, welche der betreffenden Technologiefirma gegenüber an technologischer Expertise deutlich unterlegen ist.)

Das eigenverantwortliche Trainieren des KI-Systems durch die Polizeibehörde bzw. eine andere miteinbezogene staatliche Behörde als Betreiber wird sogar – wenigstens zu einem gewissen Grad – rechtlich geboten sein. Das liegt erstens daran, dass es sich hierbei um eine wichtige begleitende oder vorbereitende Funktion für die **kriminal- oder sicherheitspolizeilicher Ermittlungen** handelt, einen besonders stark in **Grundrechte** eingreifenden Kernbereich **hoheitlicher Aufgaben**. Es wäre äußerst künstlich und nicht zweckgemäß, das Verarbeiten von Daten im Zuge des Trainierens eines KI-Systems zwecks Durchführung oder Verbesserung solcher Ermittlungen funktional von diesen abzutrennen und als nicht-hoheitliche Aufgabe anzusehen. Der Anbieter ist, sofern es sich um eine private Technologiefirma handelt, nicht zur Ausübung hoheitlicher Aufgaben, schon gar nicht im Bereich „Strafverfolgung", ermächtigt und steht außerhalb der verfassungsrechtlichen Struktur von Staatsfunktionen im Sinne der Art 18 und 20 B-VG.[81] Zweitens ist es aus Gründen des Schutzes personenbezogener Daten (insbesondere im Kontext von Art 10 Abs 5 KI-VO) sowie sensibler operativer Daten im Sinne des Art 3 Z 38 KI-VO wichtig, dass die Behörde das KI-System trainieren kann, ohne den Anbieter hinzuziehen und ihm solche Daten offenlegen zu müssen.[82] Drittens sprechen für das Trainieren durch die Behörde einfache praktische und ökonomische Gründe: „Betreiber können am besten verstehen, wie das Hochrisiko-KI-System konkret eingesetzt wird, und können somit dank einer genaueren Kenntnis des Verwendungskontextes sowie der wahrscheinlich betroffenen Personen oder Personengruppen, einschließlich schutzbedürftiger Gruppen, erhebliche potenzielle Risiken erkennen, die in der Entwicklungsphase nicht vorausgesehen wurden."[83] Es wäre womöglich nicht nur verfassungswidrig, sondern auch schlichtweg ineffizient oder kostspielig, das ermittlungsbezogene Trainieren von KI-Systemen weg von der weitaus sachnäheren Behörde auf private Akteure „auszulagern".

Auf die Gefahr der **Diskriminierung** aufgrund falscher, unvollständiger oder einseitiger Daten, die zum Trainieren des Systems verwendet werden, wurde bereits in zahlreichen Texten hingewiesen.[84] Es ist immer tunlichst darauf zu achten, dass „gesellschaftlich etablierte Vorurteile"[85] durch die un-

81 Vgl. VfSlg 14.473/1996; *Bußjäger*, "Weisungsfreie Verwaltung und Beleihung", ÖZW 2022, 190 [191].

82 ErwGr 69 KI-VO hält an zum „Einsatz von Technik, die es ermöglicht, Algorithmen direkt am Ort der Datenerzeugung einzusetzen und KI-Systeme zu trainieren, ohne dass Daten zwischen Parteien übertragen oder die Rohdaten oder strukturierten Daten selbst kopiert werden."

83 ErwGr 93 KI-VO.

84 Dazu insbesondere *Unabhängige Hochrangige Expertengruppe für künstliche Intelligenz*, Ethik-Leitlinien für eine vertrauenswürdige KI, Rz 79 ff, 111; ferner ErwGr 67 KI-VO („insbesondere in Bezug auf Personen, die bestimmten schutzbedürftigen Gruppen wie aufgrund von **Rassismus** benachteiligten oder ethnischen Gruppen angehören").

85 *Schürmann*, DSB 2021, 320 [321].

reflektierte Verwendung von KI-Systemen nicht reproduziert oder gar verhärtet, „zementiert“[86] werden. Es ist also vor dem Einsatzes des Systems und auch währenddessen laufend die nötige **Diversität** und **Ausgewogenheit** der zugrunde gelegten Trainings-Datensätze zu überprüfen. Die Funktion der Künstlichen Intelligenz darf – insbesondere im sensiblen Bereich der Strafverfolgung – nicht dazu beitragen, dass Personen beispielsweise „wegen des Geschlechts, der Rasse, der Hautfarbe, der ethnischen oder sozialen Herkunft, der genetischen Merkmale, der Sprache, der Religion oder der Weltanschauung, der politischen oder sonstigen Anschauung, der Zugehörigkeit zu einer nationalen Minderheit, des Vermögens, der Geburt, einer Behinderung, des Alters oder der sexuellen Ausrichtung“ benachteiligt werden (Art 21 GRC). Eine solche Benachteiligung liegt zweifelsohne im Falle einer verstärkten oder intensivierten behördlichen Verfolgung ohne einen sachlichen Grund vor, welcher im Verdacht einer konkret begangenen Straftat oder in einer hinreichend plausiblen Vermutung einer tatsächlich bestehenden Gefahr für bestimmte Rechtsgüter bestehen muss. In diesem Sinne sind unzureichende Trainingsdaten zugleich eine Gefahr für die **Unschuldsvermutung**.[87] Trainingsdaten bestimmen einerseits die Präzision und somit die Leistungsfähigkeit des KI-Systems, indem sie dessen Fehlerquote maßgeblich beeinflussen, andererseits wirken sie sich dadurch auf die Frage der Rechtskonformität des KI-Einsatzes aus, indem sie Ursache von faktisch unrichtigen oder diskriminierenden Ausgaben und folglich rechtswidrigen behördlichen Handelns oder Entscheidens werden können. **Trainingsdaten** stellen in diesem Sinne den „Bildungsinhalt“ des KI-Systems dar,[88] sie verkörpern dessen Kenntnis der Wirklichkeit und begründen damit seine Funktionalität für gesellschaftliche oder staatliche Aufgaben:

> „Ein diversitätssensibler Datensatz gewährleistet, dass ein Modell eine breite Palette von Eingaben korrekt verarbeiten kann. Dadurch wird das Modell wahrscheinlich in einer Vielzahl von Anwendungen und Kontexten genauer und zuverlässiger sein. […] Die Diversitätssensibilität von Trainingsdaten trägt demnach wesentlich dazu bei, die Fairness, Genauigkeit, wirtschaftliche Tragfähigkeit und ethische Integrität von KI-Systemen zu gewährleisten. Sie ist sowohl eine Frage der sozialen Verantwortung als auch der technischen Qualitätssicherung.“[89]

Sind die Trainingsdaten „nicht ausgewogen und umfassend genug“, spricht man häufig von einer **Verzerrung** der Ausgaben eines KI-Systems.[90] „[D]as mit diesen Daten trainierte System [wird]

86 Vgl. *Martini*, Blackbox Algorithmus, 57, 335.
87 Vgl. ErwGr 42, 59 KI-VO.
88 *Helminger*, EALR 2022, 46 [46].
89 *Rachinger/Voithofer/Kettemann*, jusIT 2024, 43 [49].
90 Vgl. Art 10 Abs 2 lit f und g, Art 15 Abs 4 KI-VO; *Mayrhofer/Parycek*, Digitalisierung des Rechts – Herausforderungen und Voraussetzungen, 21. ÖJT Band IV/1, 42 f.

keine angemessenen Gesetzmäßigkeiten abbilden können und möglicherweise unfaire Entscheidungen treffen, die bestimmte Gruppen gegenüber anderen bevorzugen können."[91] An ebendiesen Begriff der „Verzerrung" knüpft auch Art 10 Abs 5 KI-VO an, welcher es zum Zwecke der Korrektur und Behebung solcher Verzerrungen ausnahmsweise erlaubt, das KI-System mit sensiblen personenbezogenen Daten zu trainieren.[92] „Verzerrungen können zum Beispiel — insbesondere bei Verwendung historischer Daten — den zugrunde liegenden Datensätzen innewohnen oder bei der Implementierung der Systeme in der realen Welt generiert werden. Die von einem KI-System ausgegebenen Ergebnisse könnten durch solche inhärenten Verzerrungen beeinflusst werden, die tendenziell allmählich zunehmen und dadurch bestehende Diskriminierungen fortschreiben und verstärken".[93] Es können mithin „Rückkopplungsschleifen" entstehen, indem problematische „Datenausgaben die Eingaben für künftige Operationen beeinflussen".[94]

Produziert ein KI-System in **systematischer** Weise diskriminierende Ausgaben, kann es gar als eine „**verbotene Praktik**" im Sinne des Art 5 Abs 1 lit c KI-VO gelten. Nach dieser Verordnungsstelle ist **kategorisch verboten**:

> „das Inverkehrbringen, die Inbetriebnahme oder die Verwendung von KI-Systemen zur Bewertung oder Einstufung von natürlichen Personen oder Gruppen von Personen über einen bestimmten Zeitraum auf der Grundlage ihres sozialen Verhaltens oder bekannter, abgeleiteter oder vorhergesagter persönlicher Eigenschaften oder Persönlichkeitsmerkmale, wobei die soziale Bewertung zu einem oder beiden der folgenden Ergebnisse führt:
>
> > i) Schlechterstellung oder Benachteiligung bestimmter natürlicher Personen oder Gruppen von Personen in sozialen Zusammenhängen, die in keinem Zusammenhang zu den Umständen stehen, unter denen die Daten ursprünglich erzeugt oder erhoben wurden;
> >
> > ii) Schlechterstellung oder Benachteiligung bestimmter natürlicher Personen oder Gruppen von Personen in einer Weise, die im Hinblick auf ihr soziales Verhalten oder dessen Tragweite ungerechtfertigt oder unverhältnismäßig ist".[95]

91 *Unabhängige Hochrangige Expertengruppe für künstliche Intelligenz*, Eine Definition der KI, 6.
92 Dazu sogleich.
93 ErwGr 67 KI-VO.
94 Ebd.
95 Rechtsinformatische Notiz: Anders als beispielsweise in Art 9 und 10 KI-VO hat sich der Unionsgesetzgeber in Art 5 lobenswerter Weise der Möglichkeit bedient, die Abgrenzung der Ebenen „Absatz" und „litera" durch Einrücken der jeweiligen Textbausteine optisch hervorzuheben. Dies fördert einerseits die Übersicht für das menschliche Auge, andererseits kann es Vorteile für die maschinelle Verarbeitung von Rechtstexten mit sich bringen (zu diesem Themenkreis ausführlich *Schweighofer*, Rechtsinformatik und Wissensrepräsentation).

Wo genau die Schwelle liegt, ab welcher Diskriminierungen durch ein KI-System nicht mehr bloß als Fehler im Einzelfall, sondern als systematisch und damit hinreichend für den Tatbestand des Verbots nach Art 5 KI-VO sind, muss unter Betrachtung der technischen Spezifikationen und der faktischen Auswirkungen der KI-Verwendung im Kontext der jeweiligen Funktion geklärt werden. Es ist zu empfehlen, dass zu diesem Zweck praktisch relevante Anwendungsfälle (*use cases*) ausgearbeitet werden. Gesamthafte Erwägungen dieser Art dürfen freilich nicht darüber hinwegtäuschen, dass jede Diskriminierung als solche sowohl rechtlich als auch moralisch ein Übel darstellt, trete sie noch so singulär auf. „Der Einsatz von KI-Instrumenten durch Strafverfolgungsbehörden und anderen relevanten Behörden sollte nicht zu einem **Faktor der Ungleichheit oder Ausgrenzung** werden."[96]

Für die möglichen **Quellen, Datenarten und Methoden**, die für die Gewinnung und Verarbeitung von Trainingsdaten in Frage kommen, enthält die KI-VO zwei wesentliche Vorschriften: Zum einen verbietet Art 5 Abs 1 lit e „das Inverkehrbringen, die Inbetriebnahme für diesen spezifischen Zweck oder die Verwendung von KI-Systemen, die Datenbanken zur Gesichtserkennung durch das **ungezielte Auslesen von Gesichtsbildern aus dem Internet oder**[97] **von Überwachungsaufnahmen** erstellen oder erweitern".[98] Werden also Trainingsdaten aus den genannten Quellen bezogen, müssen diese in einem konkreten sachlichen Zusammenhang zum jeweiligen Ermittlungsakt stehen, für welchen das Trainieren durchgeführt wird. Zum anderen erlaubt Art 10 Abs 5 KI-VO – rücksichtlich der Bedingungen gemäß § 39 DSG – unter eng umschriebenen Umständen die Verarbeitung von **sensiblen personenbezogenen Daten** („besonderer Kategorien personenbezogener Daten") zu Trainings-, Test- und Validierungszwecken:

> „Soweit dies für die **Erkennung und Korrektur von Verzerrungen** im Zusammenhang mit Hochrisiko-KI-Systemen […] unbedingt erforderlich ist, dürfen die Anbieter solcher Systeme **ausnahmsweise** besondere Kategorien personenbezogener Daten verarbeiten, wobei sie angemessene Vorkehrungen für den Schutz der Grundrechte und Grundfreiheiten natürlicher Personen treffen müssen. Zusätzlich zu den Bestimmungen […] der Richtlinie (EU) 2016/680 [DSRL-PJ] müssen alle folgenden Bedingungen erfüllt sein, damit eine solche Verarbeitung stattfinden kann:

96 ErwGr 59 KI-VO.

97 Das Wort „oder" ist hier im Interesse der Grundrechte dahingehend zu verstehen, dass erstens das „ungezielte Auslesen von Gesichtsbildern aus dem Internet" und zweitens das „ungezielte Auslesen von Überwachungsmaßnahmen" unzulässig ist, hierbei nicht bloß das „ungezielte Auslesen von Gesichtsbildern von Überwachungsmaßnahmen".

98 „[...] da dies das Gefühl der Massenüberwachung verstärkt und zu schweren Verstößen gegen die Grundrechte, einschließlich des Rechts auf Privatsphäre, führen kann." (ErwGr 43 KI-VO)

a) Die Erkennung und Korrektur von Verzerrungen kann durch die Verarbeitung anderer Daten, einschließlich synthetischer oder anonymisierter Daten, **nicht effektiv durchgeführt** werden;

b) die besonderen Kategorien personenbezogener Daten unterliegen technischen Beschränkungen einer Weiterverwendung der personenbezogenen Daten und modernsten Sicherheits- und Datenschutzmaßnahmen, einschließlich Pseudonymisierung;

c) die besonderen Kategorien personenbezogener Daten unterliegen Maßnahmen, mit denen sichergestellt wird, dass die verarbeiteten personenbezogenen Daten gesichert, geschützt und Gegenstand angemessener Sicherheitsvorkehrungen sind, wozu auch strenge Kontrollen des Zugriffs und seine Dokumentation gehören, um Missbrauch zu verhindern und sicherzustellen, dass nur befugte Personen Zugang zu diesen personenbezogenen Daten mit angemessenen Vertraulichkeitspflichten haben;

d) die besonderen Kategorien personenbezogener Daten werden nicht an Dritte übermittelt oder übertragen, noch haben diese Dritten anderweitigen Zugang zu diesen Daten;

e) die besonderen Kategorien personenbezogener Daten werden **gelöscht**, sobald die Verzerrung korrigiert wurde oder das Ende der Speicherfrist für die personenbezogenen Daten erreicht ist, je nachdem, was zuerst eintritt;[99]

f) die Aufzeichnungen über Verarbeitungstätigkeiten gemäß […] der Richtlinie (EU) 2016/680 enthalten die **Gründe**, warum die Verarbeitung besonderer Kategorien personenbezogener Daten für die Erkennung und Korrektur von Verzerrungen unbedingt erforderlich war und warum dieses Ziel mit der Verarbeitung anderer Daten nicht erreicht werden konnte."

Nach der Definition des Art 9 Abs 1 DSGVO sind die gegenständlichen Daten solche, „aus denen die rassische und ethnische Herkunft, politische Meinungen, religiöse oder weltanschauliche Überzeugungen oder die Gewerkschaftszugehörigkeit hervorgehen, sowie […] genetische[] Daten, biometrische[] Daten zur eindeutigen Identifizierung einer natürlichen Person, Gesundheitsdaten oder Daten zum Sexualleben oder der sexuellen Orientierung einer natürlichen Person".[100] Die Ermächtigung dazu, solche Daten für u.a. Trainingszwecke zu verarbeiten, verfolgt nach Art 10 Abs 5 KI-VO ausdrücklich den Zweck, „Verzerrungen" im KI-System zu erkennen und zu korrigieren. Sie dient somit dem Schutz der Grundrechte, insbesondere dem Schutz vor Diskriminierung und unsachlicher behördlicher Verfolgung.

99 Zum Thema der Datenverarbeitung zu Testzwecken und zur Fehlerbehebung vgl. EuGH 20.10.2022, C-77/21 (*Digi*).
100 Siehe Art 3 Z 37 KI-VO.

3.4. Grundrechte-Folgenabschätzung

Vor der Inbetriebnahme eines Hochrisiko-KI-Systems ist zwingend eine Grundrechte-Folgenabschätzung durchzuführen. Diese Pflicht trifft gemäß Art 27 Abs 1 KI-VO alle öffentlich-rechtlichen Betreiber der in Anhang III genannten Systeme, mit Ausnahme derjenigen im Bereich „kritische Infrastruktur".[101] Somit bezieht sich diese Vorschrift zweifellos auf den Bereich Strafverfolgung einschließlich Biometrie. Es handelt sich um eine Aufgabe, die spezifisch den **Betreiber** trifft, der freilich im Sinne der obigen Ausführungen vom Anbieter dabei unterstützt werden soll.[102] Die von der Verordnung geforderte Grundrechte-Folgenabschätzung ist nach Art 27 Abs 4 ergänzend zur Datenschutz-Folgenabschätzung, somit zusätzlich zu dieser durchzuführen und umfasst nach Abs 1 jedenfalls folgende Punkte:

> „a) eine Beschreibung der **Verfahren** des Betreibers, bei denen das Hochrisiko-KI-System im Einklang mit seiner **Zweckbestimmung** verwendet wird;
> b) eine Beschreibung des **Zeitraums** und der **Häufigkeit**, innerhalb dessen bzw. mit der jedes Hochrisiko-KI-System verwendet werden soll;
> c) die **Kategorien der natürlichen Personen und Personengruppen**, die von seiner Verwendung im spezifischen Kontext **betroffen** sein könnten;
> d) die spezifischen **Schadensrisiken**, die sich auf die gemäß Buchstabe c dieses Absatzes ermittelten Kategorien natürlicher Personen oder Personengruppen auswirken könnten, unter Berücksichtigung der vom Anbieter gemäß Artikel 13 bereitgestellten Informationen;
> e) eine Beschreibung der Umsetzung von Maßnahmen der **menschlichen Aufsicht** entsprechend den Betriebsanleitungen;
> f) die **Maßnahmen**, die im Falle des Eintretens dieser Risiken zu ergreifen sind, einschließlich der Regelungen für die interne Unternehmensführung und **Beschwerdemechanismen**."

„Ziel der Grundrechte-Folgenabschätzung ist es, dass der Betreiber die spezifischen Risiken für die Rechte von Einzelpersonen oder Gruppen von Einzelpersonen, die wahrscheinlich betroffen sein werden, ermittelt und Maßnahmen ermittelt, die im Falle eines Eintretens dieser Risiken zu ergreifen sind. Die Folgenabschätzung sollte vor dem erstmaligen Einsatz des Hochrisiko-KI-Systems

101 Solche Systeme haben für die Grundrechte von Menschen ein geringeres Risiko, weil sie im Sinne des Anhang III Z 2 KI-VO lediglich als „Sicherheitsbauteile im Rahmen der Verwaltung und des Betriebs kritischer digitaler Infrastruktur, des Straßenverkehrs oder der Wasser-, Gas-, Wärme- oder Stromversorgung verwendet werden" und daher wohl nicht allzu viele personenbezogene Daten verarbeiten bzw. keine Entscheidungen in Bezug auf Personen treffen.

102 Interessanterweise ist jedoch der Begriff der **Inbetriebnahme**, vor welcher die Grundrechte-Folgenabschätzung durchzuführen ist, in Art 3 Z 11 KI-VO als eine Handlung des Anbieters definiert: „die Bereitstellung eines KI-Systems in der Union zum Erstgebrauch direkt an den Betreiber oder zum Eigengebrauch entsprechend seiner Zweckbestimmung".

durchgeführt werden, und sie sollte **aktualisiert** werden, wenn der Betreiber der Auffassung ist, dass sich einer der relevanten Faktoren geändert hat."[103]

Es wurde bereits in der Literatur auf die Wechselwirkung von Grundrechte- und **Datenschutz-Folgenabschätzung** (GRFA und DSFA) hingewiesen,[104] welche parallel zueinander laufen bzw. einander ergänzen sollen. Das Verhältnis zwischen KI-VO und Datenschutzrecht lässt sich hier folgendermaßen charakterisieren: Einerseits hat die Verwendung von KI in aller Regel Einfluss darauf, wie die Datenschutz-Folgenabschätzung zu gestalten ist, zumal der Betreiber nach Art 26 Abs 9 KI-VO gegebenenfalls die Informationen aus der KI-Betriebsanleitung dafür heranzuziehen hat; andererseits könnte es sein, dass die in Art 27 Abs 1 genannten Punkte bereits in der DSFA behandelt wurden. Für diesen Fall ordnet diese Vorschrift an, dass die GRFA die DSFA „**ergänzt**". Es ist also danach zu fragen, ob die Verwendung des Systems **KI-spezifische Risiken** birgt, die über das allgemeine Datenschutz-Risiko hinausgehen.[105] Abgrenzungsfragen können hierbei allerdings sehr schwierig werden – beispielsweise ist das automatische Schlussfolgern oder Entscheiden sowie das „Profiling" bereits vom Datenschutzrecht nach § 41 DSG („automatisierte Entscheidungsfindung im Einzelfall") erfasst –, sodass zur Gewährleistung des Grundrechtsschutzes die fraglichen Punkte im Zweifel „doppelt", nämlich sowohl aus datenschutz- als auch aus KI-rechtlicher Sicht behandelt werden sollten. Im Falle der biometrischen Fernidentifizierung, insbesondere im Rahmen von Videoaufnahmen, darf indessen angenommen werden, dass über das allgemein-datenschutzrechtliche Risiko der Überwachung hinaus das KI-spezifische Risiko der automatischen Identifikation einer Person eingeführt wird.[106]

Gemäß §§ 36 Abs 1, 52 DSG haben Strafverfolgungsbehörden eine DSFA gemäß Art 35 DSGVO durchzuführen. Für Bereiche außerhalb der Strafverfolgung hat die Datenschutzbehörde eine „Verordnung über Verarbeitungsvorgänge, für die eine Datenschutz-Folgenabschätzung durchzuführen ist", erlassen (DSFA-V).[107] Grundsätzlich ist eine DSFA nach Art 27 Abs 1 DSRL-PJ dann durchzuführen, wenn „eine Form der Verarbeitung, insbesondere bei Verwendung **neuer Technologien**, aufgrund der Art, des Umfangs, der Umstände und der Zwecke der Verarbeitung voraussichtlich ein

103 ErwGr 96 KI-VO. Dazu auch Art 27 Abs 2 KI-VO.
104 Vgl. *Schürmann*, DSB 2021, 320; *Schemmel*, CB 2024, 321 („Synergien").
105 Vgl. auch *I. Eisenberger* in *Martini/Wendehorst*, KI-VO (2024), Art 27, Rz 58.
106 So ist auch im Falle einer biometrischen Echtzeit-Fernidentifizierung in öffentlich zugänglichen Räumen zu Strafverfolgungszwecken, sofern sie überhaupt ausnahmsweise zulässig ist, im voraus eine besondere GRFA gemäß Art 5 Abs 2 KI-VO durchzuführen.
107 BGBl II 278/2018. Nach dieser Verordnung ist beispielsweise in Fällen von automatisierter Entscheidungsfindung, Profiling, der Videoüberwachung bestimmter Bereiche, des Einsatzes künstlicher Intelligenz, der Zusammenführung oder des Abgleichs von Daten aus verschiedenen Quellen zum Zweck algorithmischer Entscheidung sowie von Verarbeitungsvorgängen im höchstpersönlichen Bereich von Personen immer eine DSFA durchzuführen.

hohes Risiko für die Rechte und Freiheiten natürlicher Personen zur Folge“ hat. Ein hohes Risiko dieser Art besteht nach Art 35 Abs 3 DSGVO insbesondere in folgenden Fällen:

> „a) systematische und umfassende Bewertung persönlicher Aspekte natürlicher Personen, die sich auf automatisierte Verarbeitung einschließlich Profiling gründet und die ihrerseits als Grundlage für Entscheidungen dient, die Rechtswirkung gegenüber natürlichen Personen entfalten oder diese in ähnlich erheblicher Weise beeinträchtigen;
> b) umfangreiche Verarbeitung besonderer Kategorien von personenbezogenen Daten gemäß Artikel 9 Absatz 1 oder von personenbezogenen Daten über strafrechtliche Verurteilungen und Straftaten gemäß Artikel 10 oder
> c) systematische umfangreiche Überwachung öffentlich zugänglicher Bereiche.“

Da Art 27 Abs 1 DSRL-PJ bei der „Verwendung neuer Technologien“ zur DSFA verpflichtet, muss diese jedenfalls für KI-Systeme im Bereich der Strafverfolgung durchgeführt werden,[108] weil hier typischerweise hohe Risiken für Grundrechte, beispielsweise für die Privatsphäre gegenüber staatlichem Zugriff und für die Unschuldsvermutung bestehen.

Nach ErwGr 96 KI-VO empfiehlt es sich, in die GRFA auch „Verfahren für die Bearbeitung von **Beschwerden und Rechtsbehelfsverfahren**“ aufzunehmen, „da sie dazu beitragen könnten, Risiken für die Grundrechte in konkreten Anwendungsfällen zu mindern.“ Ein Rechtsbehelfsverfahren solcher Art schreibt die Verordnung in Art 86 vor, in Gestalt des Rechts „auf Erläuterung der Entscheidungsfindung im Einzelfall“, welches den Betreiber verpflichtet. Ferner legt es ErwGr 96 nahe, bei der Erstellung der GRFA „unabhängige Sachverständige und Organisationen der Zivilgesellschaft“ einzubeziehen.

Gemäß Art 27 Abs 3 und 5 KI-VO sind nach erfolgter GRFA deren Ergebnisse der **Marktüberwachungsbehörde** mitzuteilen. Das KI-System sollte erst nach erfolgter Mitteilung in Betrieb genommen werden.[109] Es ist auf jeden Fall ratsam, in den Prozess der Erstellung der Folgenabschätzungen die „für den Schutz der Grundrechte zuständigen Behörden“ gemäß Art 77 KI-VO einzubeziehen, welche Einsicht in die Dokumentationen nehmen und sogar technische Tests anfordern dürfen.[110]

---

108 Vgl. *Figl/Omulec* in *Knyrim,* DatKomm § 52 DSG (Stand 1.10.2023, rdb.at), Rz 10 ff; *Thiele/Wagner*, Praxiskommentar zum Datenschutzgesetz (DSG)² § 52 (Stand 1.2.2022, rdb.at), Rz 6 f.
„In Betracht kommen etwa eine Videoüberwachung in Haftanstalten oder von »Hotspots« iSd SPG, der polizeiliche Einsatz von Body-Cams, die erkennungsdienstliche Behandlung von Verdächtigen, die Rufdatenauswertung und Kommunikationsüberwachung, Zutrittskontrollsysteme, die Rasterfahndung, das Führen von Strafregistern oder die Teilnahme an Straftäterdatenbanksystemen.“ (ebd. Rz 7)

109 Anderer Ansicht *I. Eisenberger* in *Martini/Wendehorst*, KI-VO, Art 27, Rz 54.

110 Vgl. ebd., Rz 8.

3.5. Aufzeichnungspflichten/Protokollierung/Dokumentation

Gemäß Art 26 Abs 6 KI bewahren Betreiber von Hochrisiko-KI-Systemen „die von ihrem Hochrisiko-KI-System **automatisch erzeugten Protokolle**, soweit diese Protokolle ihrer **Kontrolle** unterliegen, für einen der Zweckbestimmung des Hochrisiko-KI-Systems angemessenen Zeitraum von mindestens **sechs Monaten** auf, sofern im geltenden Unionsrecht, insbesondere im Unionsrecht über den Schutz personenbezogener Daten, oder im geltenden nationalen Recht nichts anderes bestimmt ist."

Allerdings sind die Pflichten im Zusammenhang mit den „automatischen erzeugten Protokollen" gemäß Art 12 und deren Aufbewahrung nach Art 19 KI-VO insbesondere dem **Anbieter** zugewiesen. Dieser ist auch für die **technische Dokumentation** im Sinne des Art 11 und Anhang IV zuständig, welche primär als Unterlage für das Konformitätsbewertungsverfahren dient. Die Protokolle, welche gemäß Art 12 automatisch erzeugt und vom Betreiber nach Art 26 Abs 6 aufbewahrt werden, dienen jedoch vor allem der „**Rückverfolgbarkeit** der konkreten Ereignisse der KI-Systeme ex post".[111] Dadurch soll das „Blackbox"-Risiko reduziert werden.[112] Art 12 Abs 1 gibt vor: „Die Technik der Hochrisiko-KI-Systeme muss die automatische Aufzeichnung von Ereignissen (im Folgenden „Protokollierung") während des **Lebenszyklus** des Systems ermöglichen." Dabei liegt es nahe, dass der Anbieter zur Gewährleistung dieser Protokollierung von technischen Normen nach Art 40 Gebrauch macht. Es geht nach Art 12 Abs 2 um die „Aufzeichnung von **Ereignissen**, die für Folgendes relevant sind:

> a) die Ermittlung von Situationen, die dazu führen können, dass das Hochrisiko-KI-System ein Risiko im Sinne des Artikels 79 Absatz 1 birgt oder dass es zu einer wesentlichen Änderung kommt,
> b) die Erleichterung der Beobachtung nach dem Inverkehrbringen gemäß Artikel 72 und
> c) die Überwachung des Betriebs der Hochrisiko-KI-Systeme gemäß Artikel 26 Absatz 5."

Was genau die angeführten Artikel beinhalten, wird weiter unten im nächsten Abschnitt über die Zusammenarbeit des Betreibers mit dem Anbieter und mit der Marktüberwachungsbehörde erläutert. Ferner dürfte hier dem Begriff des „schwerwiegenden Vorfalls" im Sinne des Art 73 KI-VO Bedeutung zukommen.

111 *Braun Binder/Egli* in *Martini/Wendehorst*, KI-VO, Art 12, Rz 5. Dort mit Hervorhebungen.
112 Vgl. ebd., Rz 1 f.

Das „Funktionieren des Hochrisiko-KI-Systems“ muss nach Art 12 Abs 2 KI-VO „in einem der **Zweckbestimmung** des Systems angemessenen Maße rückverfolgbar“ sein.[113] Für den Bereich „Strafverfolgung“ bedeutet das, dass die Protokollierung möglichst detailliert und sorgfältig stattfinden muss, weil in besonderem Ausmaß das Risiko weitreichender Grundrechtseingriffe besteht und, zumal wenn es zu einem Gerichtsverfahren kommt, oftmals Beweisfragen auftreten.[114] Darüber hinaus sind solche Ereignisse automatisch aufzuzeichnen, welche im Zusammenhang mit einer „wesentlichen Veränderung“ am KI-System nach Art 3 Z 23 stehen könnten. Dort wird dieser Begriff definiert als „eine Veränderung eines KI-Systems nach dessen Inverkehrbringen oder Inbetriebnahme, die in der vom Anbieter durchgeführten ursprünglichen Konformitätsbewertung nicht vorgesehen oder geplant war und durch die die Konformität des KI-Systems mit den Anforderungen in Kapitel III Abschnitt 2 beeinträchtigt wird oder die zu einer Änderung der Zweckbestimmung führt, für die das KI-System bewertet wurde“. Unter Umständen kann das Trainieren des KI-Systems durch den Betreiber wie bereits besprochen eine solche wesentliche Veränderung darstellen, insbesondere wenn nicht im voraus der Umfang dieses Trainierens hinreichend klar und präzise zwischen dem Anbieter und dem Betreiber festgelegt wurde.

Art 12 Abs 3 KI-VO enthält eine Sondervorschrift für Systeme der **biometrischen Fernidentifizierung**, nach welcher die automatische Protokollierung mindestens den Zeitraum der Verwendung des Systems, die Referenzdatenbank für den Abgleich, die Eingabedaten und die Identität der Organe für die besondere menschliche Aufsicht nach Art 14 Abs 5 umfasst.

Art 19 Abs 1 KI-VO regelt die Aufbewahrung der Protokolle durch die **Anbieter**, „soweit diese Protokolle ihrer Kontrolle unterliegen.“ Gleichermaßen regelt Art 26 Abs 6 die Aufbewahrung der Protokolle durch die **Betreiber**, „soweit diese Protokolle ihrer Kontrolle unterliegen“. Demnach stellt sich die Frage, wer denn eigentlich in welchem Umfang Kontrolle über die Protokolle hat. Dies kann einerseits vom Anbieter in der Betriebsanleitung festgelegt werden, andererseits ist es denkbar, diese Frage staatlich durch Gesetz oder Verordnung zu regeln; in diesem Kontext wird zwischen faktischer und rechtlicher Kontrolle unterschieden, welche bestenfalls eine Einheit bildet.[115] Die tatsächliche Aufbewahrungsfrist hat sich jedenfalls nach der Zweckbestimmung des KI-Systems zu richten und mindestens sechs Monate zu betragen, „sofern im geltenden Unionsrecht, insbesondere im Unionsrecht über den Schutz personenbezogener Daten, oder im geltenden nationalen Recht nichts anderes bestimmt ist.“ Beschränkungen der Aufbewahrungszeit dürfen zum Schutz der Grundrechte, insbesondere für den Datenschutz, von der Gesetzgebung des Mitgliedsstaats vorgese-

113 Zum Begriff der „Zweckbestimmung“ siehe Art 3 Z 12 KI-VO.
114 Vgl. *Braun Binder/Egli* in *Martini/Wendehorst*, KI-VO, Art 12, Rz 16.
115 Vgl. *I. Eisenberger* in *Martini/Wendehorst*, KI-VO, Art 26, Rz 44; *dies.*, ebd., Art 19, Rz 17.

hen werden.[116] Es stellt sich allerdings die Frage, ob nicht in bestimmten Fällen gerade eine längere Aufbewahrung aus Beweisgründen eher im Interesse der Grundrechte betroffener Personen läge. Dies setzt wiederum voraus, dass die betroffene Person auf die Protokolle Zugriff hat oder sie anderweitig in einem rechtsförmigen Verfahren, etwa durch die Möglichkeit einer Überprüfung des Protokolls durch ein Gericht, geltend machen kann.

3.6. Zusammenarbeit mit dem Anbieter und der Marktüberwachungsbehörde

Die Pflichten, welche die KI-VO Anbietern von KI-Systemen auferlegt, sind vielfältig sowie teilweise sehr komplex und können hier nicht alle im Detail behandelt werden. Das betrifft auch das Zusammenspiel zwischen dem Anbieter und den involvierten Behörden. Die Verantwortung des Betreibers innerhalb dieses Gefüges ist hingegen relativ überschaubar. Die zentrale Stelle dafür ist Art 26 Abs 5, welcher zunächst einmal eine **Überwachungspflicht** für den Betreiber eines Hochrisiko-KI-Systems vorsieht:

„Die Betreiber überwachen den Betrieb des Hochrisiko-KI-Systems anhand der Betriebsanleitung und informieren gegebenenfalls die Anbieter gemäß Artikel 72." Dieser regelt die „**Beobachtung nach dem Inverkehrbringen** durch die Anbieter und Plan für die Beobachtung nach dem Inverkehrbringen für Hochrisiko-KI-Systeme". Nach Art 3 Z 25 KI-VO umfasst das System zur Beobachtung nach dem Inverkehrbringen „alle Tätigkeiten, die Anbieter von KI-Systemen zur **Sammlung und Überprüfung von Erfahrungen** mit der Verwendung der von ihnen in Verkehr gebrachten oder in Betrieb genommenen KI-Systeme durchführen, um festzustellen, ob unverzüglich nötige **Korrektur- oder Präventivmaßnahmen** zu ergreifen sind". Gleichwohl dieser Begriff unzweideutig auf den Anbieter bezogen ist, hat er doch eine nicht zu vernachlässigende Bedeutung für den Betreiber. Laut Art 72 Abs 2 nämlich „müssen sich die einschlägigen Daten zur Leistung[117] der Hochrisiko-KI-Systeme, die von den Anbietern oder den Betreibern bereitgestellt oder aus anderen Quellen erhoben werden können, über ihre gesamte Lebensdauer hinweg aktiv und systematisch erheben, dokumentieren und analysieren lassen". Der Betreiber trägt somit, insofern er einschlägige Daten erhebt oder auf sonstige Weise verarbeitet, eine gewisse Mitverantwortung an der Beobachtung des Systems. Jedoch sind „sensible operative Daten"[118] der Strafverfolgungsbehörden von der Pflicht zur Offenlegung ausgenommen.

116 Vgl. ebd., Rz 16. Für einen **Vorrang datenschutzrechtlicher Bestimmungen**, insbesondere hinsichtlich des Rechts auf Löschung, *Braun Binder/Egli* ebd., Art 12, Rz 11.

117 Der Begriff der „Leistung" wird in Art 3 Z 18 KI-VO, wenig überraschend, definiert als „die Fähigkeit eines KI-Systems, seine Zweckbestimmung zu erfüllen" (vgl. Z 12).

118 Art 3 Z 38 KI-VO.

Weiter nimmt Art 26 Abs 5 KI-VO die Betreiber in die Pflicht, jede Vermutung eines Risikos nach Art 79, das heißt eines Risikos „für die Gesundheit oder Sicherheit oder Grundrechte von Personen“,[119] nicht nur dem Anbieter, sondern auch der Marktüberwachungsbehörde unverzüglich zu melden.[120] Von mindestens genauso großer Tragweite ist die Verpflichtung, in solchen Fällen überhaupt den **Betrieb des Systems auszusetzen**. Beide Maßnahmen haben ohne Zögern „ab Kenntnis der Risiken oder ab Kenntnis der Gründe, die ein Risiko vermuten lassen“,[121] zu erfolgen. Den Anbieter trifft gemäß Art 20 Abs 1 die Pflicht, das System bei mangelnder Konformität mit der KI-VO im Notfall „zurückzunehmen, zu **deaktivieren** oder zurückzurufen“,[122] falls andere Korrekturmaßnahmen nicht möglich oder nicht zielführend sind, und den Betreiber zu informieren. Wird sich der Anbieter eines Risikos im Sinne des Art 79 bewusst, „führt er unverzüglich gegebenenfalls gemeinsam mit dem meldenden Betreiber eine Untersuchung der Ursachen durch und informiert er die Marktüberwachungsbehörden“.[123]

Schließlich gelten nach Art 26 Abs 5 KI-VO besondere Regeln für den Betreiber im Zusammenhang mit „**schwerwiegenden Vorfällen**“. Nach der Definition in Art 3 Z 49 ist darunter zu verstehen ein „Vorfall oder eine Fehlfunktion bezüglich eines KI-Systems, das bzw. die direkt oder indirekt eine der nachstehenden Folgen hat:

> a) den Tod oder die schwere gesundheitliche Schädigung einer Person;
> b) eine schwere und unumkehrbare Störung der Verwaltung oder des Betriebs **kritischer Infrastrukturen**;
> c) die Verletzung von Pflichten aus den Unionsrechtsvorschriften zum Schutz der **Grundrechte**;
> d) schwere Sach- oder Umweltschäden“.

Es handelt sich demnach um einen tendenziell weiten Begriff: Nach lit c liegt ein solcher schwerwiegender Vorfall insbesondere vor, wenn eine Person aufgrund einer fehlerhaften Ausgabe des KI-Systems festgenommen wird (Art 6 GRC), ihre Wohnung oder sonstige Räumlichkeiten behördlich durchsucht werden, ihr Briefverkehr oder ihre Telekommunikation – mit welchen Mitteln auch immer – überwacht wird oder sonstige bedeutsame Eingriffe in ihren privaten Lebensbereich stattfin-

119 Art 79 Abs 1 KI-VO verweist dabei auf Art 3 Z 19 der Verordnung (EU) 2019/1020 über Marktüberwachung und die Konformität von Produkten.

120 „Haben Betreiber Grund zu der Annahme, dass die Verwendung gemäß der Betriebsanleitung dazu führen kann, dass dieses Hochrisiko-KI-System ein Risiko im Sinne des Artikels 79 Absatz 1 birgt, so informieren sie unverzüglich den Anbieter oder Händler und die zuständige Marktüberwachungsbehörde […].“ (Art 26 Abs 5 KI-VO)

121 *I. Eisenberger* in *Martini/Wendehorst*, KI-VO, Art 26, Rz 37.

122 Dazu Art 3 Z 16 und 17 KI-VO.

123 Art 20 Abs 2 KI-VO. Die Zusammenarbeit des Anbieters mit den Behörden ist ferner in Art 21 geregelt.

den (Art 7 und 8 GRC).[124] Andererseits umfasst lit b der zitierten Bestimmung **Cyber-Angriffe** auf staatliche Infrastruktur, beispielsweise solche der öffentlichen Sicherheit;[125] eine „unumkehrbare" Störung derselben liegt wohl dann vor, wenn sie „nicht im Wege üblicher Instandhaltungsarbeiten, für die jederzeit vorgesorgt sein muss, beseitigt werden kann."[126] Generell können auch „Bedien- bzw. Anwendungsfehler" einen schwerwiegenden Vorfall darstellen.[127]

Ein solcher Vorfall ist gleichermaßen dem Anbieter und der Marktüberwachungsbehörde zu melden. Die **Marktüberwachungsbehörde (MÜB)**[128] ist im Bereich der Strafverfolgung nach der Sondervorschrift des Art 74 Abs 8 KI-VO in Verbindung mit §§ 31, 36 DSG grundsätzlich die **Datenschutzbehörde**. Allerdings erlaubt Art 74 Abs 8 KI-VO auch eine „andere gemäß denselben Bedingungen wie den in den Artikeln 41 bis 44 der Richtlinie (EU) 2016/680 festgelegten benannte Behörde". Der MÜB „obliegt der größte Teil der Durchsetzung der KI-VO",[129] sie darf jedoch in „keiner Weise die Unabhängigkeit von Justizbehörden beeinträchtigen oder deren Handlungen im Rahmen ihrer justiziellen Tätigkeit anderweitig beeinflussen."[130] Sie ist ferner im Spezialfall der biometrischen KI-Systeme der Strafverfolgungsbehörden nach Art 43 Abs 1 KI-VO für das Konformitätsbewertungsverfahren zuständig, welches jedoch in erster Linie den Anbieter, nicht so sehr den Betreiber betrifft.[131]

Die Meldung schwerwiegender Vorfälle wird ist erster Linie in Art 73 KI-VO geregelt. Art 26 Abs 5, welcher wiederum eine Ausnahme für sensible operative Daten enthält, normiert speziell für Betreiber: „Haben die Betreiber einen schwerwiegenden Vorfall festgestellt, informieren sie auch unverzüglich zuerst den Anbieter und […] die zuständigen Marktüberwachungsbehörden über diesen Vorfall. **Kann der Betreiber den Anbieter nicht erreichen**, so gilt Artikel 73 entsprechend."[132] Hat der Anbieter, beispielsweise aufgrund einer Meldung des Betreibers oder auch aus eigener Wahr-

124 Vgl. *Wendehorst* in *Martini/Wendehorst*, KI-VO, Art 3, Rz 336.
125 Dazu Art 3 Z 62 KI-VO; Art 2 Z 4 und 5 RL (EU) 2022/2557.
126 *Wendehorst* in *Martini/Wendehorst*, KI-VO, Art 3, Rz 335.
127 *Hartmann* in *Martini/Wendehorst*, KI-VO, Art 73, Rz 7. Dort hervorgehoben.
128 Siehe Art 3 Z 26 und Art 74 Abs 1 KI-VO mit Verweis auf VO 2019/1020, vgl. dort insbesondere Art 11, 14 und 16 zu den Aufgaben und Befugnissen der Behörde.
129 *Wendehorst* in *Martini/Wendehorst*, KI-VO, Art 3, Rz 203. Dort mit Hervorhebung.
130 Art 74 Abs 8 KI-VO.
131 Für sonstige KI-Systeme im Bereich Strafverfolgung sieht Art 43 Abs 2 KI-VO ein „Konformitätsbewertungsverfahren auf der Grundlage einer internen Kontrolle gemäß Anhang VI" ohne Beteiligung einer notifizierten Stelle vor. Jedoch darf der Anbieter wohl auch freiwillig die externe Bewertung nach Anhang VII unter Beiziehung einer notifizieren Stellen wählen (vgl. *Gerdemann* in *Martini/Wendehorst*, KI-VO, Art 43, Rz 35, 21). Zu den möglichen Ausnahmen vom Konformitätsbewertungsverfahren siehe Art 46, dessen Abs 2 Strafverfolgungsbehörden sogar eine Inbetriebnahme mit bloß nachträglicher Genehmigung in „hinreichend begründeten dingenden Fällen aus außergewöhnlichen Gründen der öffentlichen Sicherheit oder in Fällen einer konkreten, erheblichen und unmittelbaren Gefahr für das Leben oder die körperliche Unversehrtheit natürlicher Personen" erlaubt.
132 Der letzte Satz bedeutet, dass der Betreiber in solchen Fällen die Pflichten des Anbieters nach Art 73 KI-VO im Rahmen seiner Möglichkeiten selbst erfüllen muss (nach *I. Eisenberger* in *Martini/Wendehorst*, KI-VO, Art 26, Rz 38).

nehmung, von einem „schwerwiegenden Vorfall“ erfahren, muss er dies nach Art 73 Abs 1, 2, 3 und 4 unverzüglich der MÜB melden. Genauer hat diese Meldung unmittelbar zu erfolgen, „nachdem der Anbieter **kausalen Zusammenhang** zwischen dem KI-System und dem schwerwiegenden Vorfall oder die **naheliegende Wahrscheinlichkeit** eines solchen Zusammenhangs festgestellt hat und in jedem Fall spätestens **15 Tage**, nachdem der Anbieter oder gegebenenfalls der Betreiber Kenntnis von diesem schwerwiegenden Vorfall erlangt hat.“ Bezüglich dieses „Meldezeitraums wird der Schwere des schwerwiegenden Vorfalls Rechnung getragen.“ Besteht der schwerwiegende Vorfall in einer Störung kritischer Infrastruktur nach Art 3 Z 49, verkürzt sich der maximale Zeitraum auf zwei Tage ab Kenntnis des Vorfalls.[133] Die Pflicht zur Meldung ist gemäß Art 73 Abs 5 KI-VO als eine **Berichtspflicht** konzipiert: „Wenn es zur Gewährleistung der rechtzeitigen Meldung erforderlich ist, kann der Anbieter oder gegebenenfalls der Betreiber einen **unvollständigen Erstbericht** vorlegen, dem ein vollständiger Bericht folgt.“ Laut Abs 6 muss der Anbieter im Anschluss an die Meldung „unverzüglich die erforderlichen **Untersuchungen** im Zusammenhang mit dem schwerwiegenden Vorfall und dem betroffenen KI-System“ durchführen. „Dies umfasst eine Risikobewertung des Vorfalls sowie Korrekturmaßnahmen.“ Dabei hat er mit den Behörden zusammenzuarbeiten, wobei er ohne Beiziehung der MÜB keine Veränderung am KI-System vornehmen darf, „die möglicherweise Auswirkungen auf eine spätere Bewertung der Ursachen des Vorfalls hat“. Die MÜB informiert wiederum die für den Schutz der Grundrechte zuständigen Behörden nach Art 77,[134] ergreift geeignete Maßnahmen zur Gewährleistung der Produktsicherheit – welche gegebenenfalls die Untersagung des KI-Systems umfassen –[135] und meldet den Vorfall der Europäischen Kommission.[136]

Aufgaben und Befugnisse der MÜB regelt die KI-VO vor allem in Art 74, 79, 80, 82 und 83. Auf die besonders zahlreichen Details kann hier nicht näher eingegangen werden; jedoch ist festzuhalten, dass einige Befugnisse sich nicht bloß gegen den Anbieter, sondern gegen jeden „Akteur“ richten, welcher gemäß Art 3 Z 8 auch der Betreiber sein kann. Darüber hinaus verlangt Art 77 von den Mitgliedsstaaten, **für den Schutz der Grundrechte zuständige Behörden** zu benennen,[137] welche bei Hochrisiko-KI-Systemen nach Anhang III Dokumentationen anfordern und einsehen dürfen sowie gegebenenfalls technische Tests bei der MÜB beantragen können.

133 Ferner enthält Art 73 Abs 4 KI-VO eine besondere Frist für den Fall des Todes einer Person.
134 Siehe Art 73 Abs 7 KI-VO.
135 Siehe Art 73 Abs 8 KI-VO und Art 19 VO (EU) 2019/1020.
136 Siehe Art 73 Abs 11 KI-VO.
137 Dies muss nach Art 77 Abs 2 KI-VO schon im November 2024 geschehen sein.

Art 78 KI-VO verpflichtet alle Personen und Stellen, „die an der Anwendung dieser Verordnung beteiligt sind“, zu einem vertraulichen Umgang mit allen Informationen und Daten, „in deren Besitz sie bei der Ausführung ihrer Aufgaben und Tätigkeiten gelangen“. Dabei werden „öffentliche und nationale Sicherheitsinteressen“ und „die Durchführung von Straf- oder Verwaltungsverfahren“ eigens genannt, Abs 2 hält Behörden zu Maßnahmen der Cybersicherheit an. Geheimhaltungs- und Vertraulichkeitspflichten nach dem Datenschutzrecht sowie der Grundsatz der Datenminimierung gelten indessen unverändert auch im Anwendungsbereich der KI-VO.[138] Vertrauliche Informationen von Strafverfolgungsbehörden dürfen laut Abs 3 „nicht ohne vorherige Rücksprache mit der zuständigen nationalen Behörde, von der die Informationen stammen, und dem Betreiber offengelegt werden, sofern […] eine solche Offenlegung die öffentlichen und nationalen Sicherheitsinteressen gefährden könnte. Dieser Informationsaustausch erstreckt sich nicht auf sensible operative Daten“.[139]

Hochrisiko-KI-Systeme nach Anhang III[140] sind vom Anbieter vor dem Inverkehrbringen oder der Inbetriebnahme in eine **EU-Datenbank** nach Art 49 und 71 KI-VO mit den in Anhang VIII und IX geforderten Informationen einzutragen. Agiert eine **Behörde als Betreiber** eines solchen KI-Systems, hat sie nach Art 49 Abs 3 vor der Inbetriebnahme oder der Verwendung ebenfalls eine Registrierung desselben vorzunehmen. Dies dient der Arbeitserleichterung und der „Transparenz gegenüber der Öffentlichkeit“.[141] Für KI-Systeme im Bereich der Strafverfolgung gilt allerdings gemäß Abs 4 die Besonderheit, dass sie in einem nicht öffentlichen Teil der Datenbank registriert werden, zu dem nur die Kommission und die zuständige MÜB Zugang haben, und dass beispielsweise „die Beschreibung der verwendeten Daten und Inputs und der Betriebslogik des KI-Systems […] und die elektronische Betriebsanleitung“, ferner die Grundrechte- und die Datenschutz-Folgenabschätzung überhaupt von der Registrierungspflicht ausgenommen sind.[142]

138 Zudem bestimmt Art 78 Abs 4 KI-VO: „Die Absätze 1, 2 und 3 dürfen sich weder auf die Rechte oder Pflichten der Kommission, der Mitgliedstaaten und ihrer einschlägigen Behörden sowie der notifizierten Stellen in Bezug auf den Informationsaustausch und die Weitergabe von Warnungen, einschließlich im Rahmen der grenzüberschreitenden Zusammenarbeit, noch auf die Pflichten der betreffenden Parteien auswirken, Informationen auf der Grundlage des Strafrechts der Mitgliedstaaten bereitzustellen.“

139 Übernimmt die Strafverfolgungsbehörde zugleich die Rolle des Anbieters, so hat die technische Dokumentation des KI-Systems in ihren Räumlichkeiten zu bleiben und das Personal der MÜB muss vor der Einsichtnahme eine „Sicherheitsfreigabe“ durchlaufen.
Vgl. parallel dazu § 12 Abs 4 SNG.

140 Im Sinne des Art 6 Abs 3 KI-VO als nicht hochriskant eingestufte Systeme sind nach Art 49 Abs 2 ebenfalls einzutragen, wobei im Bereich Strafverfolgung gemäß Abs 4 lit b die rechtlichen Bedingungen und faktischen Gründe für diese Einstufung ausgenommen sind. Gemäß Art 49 Abs 5 werden KI-Systeme der „kritischen Infrastruktur“ auf nationaler Ebene registriert.

141 ErwGr 131 KI-VO.

142 *Gerdemann* in *Martini/Wendehorst*, KI-VO, Art 49, Rz 31, 33, mit Nachweisen. Dort mit Hervorhebungen.

4. KI-Einsatz in Ermittlungen nach StPO, SPG, SNG

Die bisherigen Ausführungen waren vor allem der Darstellung der Betreiberpflichten nach der KI-VO gewidmet, welche oftmals eng miteinander verwoben sind. In den nächsten Abschnitten geht es um die Frage, welche Bedeutung der KI-VO für konkrete strafprozessuale oder sicherheitspolizeiliche Ermittlungen zukommt, wobei hier die biometrischen Fernidentifizierung – nachträglich oder in Echtzeit – und bestimmte Informationspflichten gegenüber der betroffenen Personen zentral sind. Im Übrigen gibt es eine breite Menge an möglichen Technologien und KI-unterstützten Verfahren, die in solchen Ermittlungen zum Einsatz kommen könnten, aber von der KI-VO nicht näher im Detail geregelt werden, sondern nur allgemein als KI-Systeme im Bereich „Strafverfolgung" nach Anhang III Z 6.

Für **KI-Systeme, welche bereits von Behörden verwendet werden**, setzt Art 111 Abs 2 KI-VO den 2. August **2030** als das Ende der **Umsetzungsfrist** fest, bis zu welchem alle „erforderlichen Maßnahmen für die Erfüllung der Anforderungen und Pflichten dieser Verordnung" gesetzt sein müssen. Allerdings gilt gemäß Art 113 lit a das Verbot von Praktiken nach Art 5 bereits ab 2. Februar **2025**. Das bedeutet, dass schon ab diesem Datum beispielsweise diskriminierende KI-Praktiken, „social scoring", ausschließlich oder vorwiegend[143] auf KI gestütztes „predictive policing" usw. einen Verstoß gegen die Verordnung begründen. Laut Art 113 lit b können ab 2. August 2025 auch **Geldbußen** dafür verhängt werden.

Von der Anwendbarkeit der KI-VO streng zu unterscheiden ist jedoch die Frage, ob das **nationale Recht** bestimmte KI-Anwendungen durch Behörden zulässt. Dass die KI-VO einen Rahmen vorgibt, innerhalb dessen ein Mitgliedsstaat KI-Systeme einsetzen darf, lässt noch keineswegs darauf schließen, dass die Behörden dieses Mitgliedsstaats sie tatsächlich einsetzen dürfen. Nach der österreichischen Verfassung bedarf hoheitliches Handeln, zumindest wenn es potenziell in Grundrechte eingreift, stets einer gesetzlichen Grundlage. Beispielsweise bestimmt § 1 Abs 2 DSG als Verfassungsbestimmung, dass behördliche Eingriffe in das Recht auf Datenschutz nur aufgrund von Gesetzen erfolgen dürfen, welche im Sinne der EMRK verhältnismäßig sind und angemessene Garantien für Betroffene bieten; zudem müssen sie dem **Bestimmtheitsgebot** des Art 18 B-VG[144] genügen und mit der europäischen Grundrechte-Charta in der Rechtsprechung des EuGH im Einklang stehen. Speziell im Bereich des **Verfassungsschutzes** hat der VfGH judiziert, dass selbst eine Recherche in öffentlich zugänglichen Internetquellen,[145] wie sie in § 10 Abs 5 SNG und § 53 Abs 4 SPG

143 Vgl. *Wendehorst* in *Martini/Wendehorst*, KI-VO, Art 5, Rz 84.
144 Dies gilt sowohl für die Verwaltung als auch für die Justiz: *Muzak,* B-VG⁶ Art 18 (Stand 1.10.2020, rdb.at), Rz 3.
145 Dazu näher im Detail mit kritischer Auseinandersetzung: *Heißl* in *Heißl/Figl* (Hrsg), SNG², § 10, Rz 35-45.

vorgesehen ist, darunter auch der bloße Gebrauch von Suchmaschinen, für Behörden keine Selbstverständlichkeit darstellt, sondern vielmehr eine gesetzliche Ermächtigung voraussetzt:

> „Entscheidet sich die Vollziehung dazu, in systematischer Weise auch öffentlich zugängliche Quellen zu nutzen, bedarf es einer ausdrücklichen gesetzlichen Grundlage, die in verfassungsrechtlich unbedenklicher Weise dazu ermächtigt. Denn auch die Ermittlung und Weiterverarbeitung öffentlich zugänglicher personenbezogener Daten unterliegt dem verfassungsgesetzlich gewährleisteten Auskunftsanspruch gemäß §1 Abs 3 DSG 2000, zumal für den Einzelnen mitunter kaum noch nachvollziehbar ist, welche Daten über ihn öffentlich verfügbar sind.“[146]

Demnach bedürfen polizeiliche Techniken wie zum Beispiel *social network analysis* und biometrische Verfahren, um überhaupt zulässig zu sein, einer gesetzlichen Grundlage. Diese fehlt jedoch nach geltendem Recht. Wenn schon der bloße Gebrauch von Suchmaschinen im Gesetz Deckung finden muss, dann erst recht die Ermittlung, Analyse und Gliederung personenbezogener Daten mittels KI-Systemen, insbesondere wenn dabei Profile über Personen oder Gruppen von Personen erstellt werden. Das liegt daran, dass die Eingriffsintensität dadurch erhöht wird:[147]

> „Auch bei der Recherche im Internet (OSINT-Recherchen) kann KI unterstützen. Zu bedenken ist allerdings, dass es idR um den Umgang mit personenbezogenen Daten geht. Wenn also etwa ein KI-Rechercheprogramm das Internet systematisch nach relevanten Informationen zum Beschuldigten durchforsten und Vernetzungen iSv Persönlichkeitsprofilen herstellen soll, dann ändert sich die Qualität der damit durchgeführten KI-Tätigkeit. Aus der bloßen Hilfe zur Effizienzsteigerung wird uU ein eigenständiger Eingriff in das Grundrecht auf Datenschutz; diesfalls bedürfte es naturgemäß einer ausdrücklichen Rechtsgrundlage für diese Art der Datenverarbeitung.“[148]

Ganz in diesem Sinne spricht auch die KI-VO an bestimmten Stellen, beispielsweise in Anhang III Z 1 und 6 über Biometrie und Strafverfolgung, von KI-Systemen, „soweit ihr Einsatz nach einschlägigem Unionsrecht oder nationalem Recht zugelassen ist“. Folglich kann der Einsatz dieser Systeme nicht direkt auf die KI-VO gestützt werden, sondern er bedarf einer eigenen, speziellen Rechtsgrundlage.

146 VfGH 29.11.2017, G 223/2016, VfSlg 20.213/2016.
147 Vgl. *Heißl* in *Heißl/Figl* (Hrsg), SNG², § 10, Rz 47.
148 *Reindl-Krauskopf*, Grundrechtliche und strafprozessrechtliche Grenzen: Welche rote Linien zieht das Recht?, AnwBl 2024, 35 [39], mwN.

Arbeitet das KI-System jedoch nicht mit personenbezogenen Daten und lassen dessen Ausgaben keine Rückschlüsse auf bestimmte Personen zu, liegt also insofern kein Eingriff in Grundrechte vor, werden an den KI-Einsatz weniger strenge Anforderungen zu stellen sein, beispielsweise im Wissensmanagement, bei der Übersetzung von Texten, bei der Arbeit mit anonymen, statistischen Daten, bei der Anonymisierung von Daten,[149] womöglich auch in der Öffentlichkeitsarbeit. Es ist immer darauf abzustellen, ob personenbezogene Daten verarbeitet oder konkrete Ermittlungen gegen Personen geführt werden. Die KI-unterstützte Erkennung von Gegenständen oder Symbolen ist jedenfalls weniger streng zu beurteilen als die von Personen, letztere wird in vielen Fällen den besonderen Vorschriften für biometrische Systeme nach Anhang III Z 1 KI-VO unterliegen. Sollen bei der jüngst wieder eingeführten, neu formulierten „section control"[150] nach § 54 Abs 4b SPG[151] Bildaufnahmen von Fahrzeugen mittels KI ausgewertet werden, muss dafür eine gesetzliche Ermächtigung bereitgestellt werden, zumal die Fahrzeugdaten Schlüsse auf Personen ermöglichen. Beim Begriff der *sentiment analysis* ist zu differenzieren: Einerseits könnte es sich dabei um eine **Emotionserkennung**, sonach um ein biometrisches Hochrisiko-KI-System nach Anhang III Z 1 lit c KI-VO handeln, welches im Sinne des Art 3 Z 39 „dem Zweck dient, Emotionen oder Absichten natürlicher Personen auf der Grundlage ihrer biometrischen Daten festzustellen oder daraus abzuleiten". Das Bestehen einer wissenschaftlichen Grundlage für ein solches Vorgehen bzw. dessen tatsächliche Zuverlässigkeit darf durchaus angezweifelt werden.[152] – Fest steht jedoch, dass es auch nach nationalem Recht besonders streng zu beurteilen wäre. Andererseits könnte es lediglich darum gehen, unter Auswertung anonymisierter[153] Daten im Rahmen der **Trendforschung** ein abstraktes **Stimmungsbild** zu erstellen, beispielsweise im Zuge der Beobachtung eines Internetforums.[154] Allerdings dürfe auch in diesem Fall, falls ein näher abgrenzbarer Personenkreis beobachtet wird, von einem gewissen Grundrechtseingriff auszugehen sein.

Je schwerer der Grundrechtseingriff, desto strenger und präziser muss das Gesetz die Behörden binden.[155] Daher ist die KI-basierte Erkennung von **Gesichtsbildern oder Stimmen** nicht von § 10 Abs 4 SNG gedeckt, auch dürfen nach § 11 Abs 1 ermittelte Standort- oder Reisedaten nicht mittels KI

149 Vgl. § 54 Abs 4b SPG idF BGBl. I Nr. 122/2024: „Soweit die bildgebende Erfassung von Personen technisch nicht ausgeschlossen werden kann, sind diese Personen ohne unnötigen Verzug in nicht rückführbarer Weise unkenntlich zu machen." Siehe auch ErwGr 61 KI-VO.

150 Zur Aufhebung der Vorgängerversion wegen Grundrechtsverletzung: VfGH 11.12.2019, G 72/2019 ua.

151 Fassung BGBl. I Nr. 122/2024.

152 Vgl. *Ruschemeier* in *Martini/Wendehorst*, KI-VO, Anhang III, Rz 20, mwN; ErwGr 44 KI-VO.

153 Zu dieser Thematik *Wendehorst* in *Martini/Wendehorst*, KI-VO, Art 3, Rz 237.

154 In die Richtung der „Trendforschung" scheint auch ErwGr 42 zu gehen, welcher von KI-Systemen spricht, „die Risikoanalysen zur Bewertung der Wahrscheinlichkeit eines Finanzbetrugs durch Unternehmen auf der Grundlage **verdächtiger Transaktionen** durchführen, oder Risikoanalyseinstrumente einsetzen, um die Wahrscheinlichkeit vorherzusagen, dass Betäubungsmittel oder illegale Waren durch Zollbehörden, beispielsweise auf der Grundlage bekannter **Schmuggelrouten**, aufgespürt werden." Dies soll nicht vom Verbot des „predictive policing" erfasst sein.

155 Vgl. *Heißl* in *Heißl/Figl* (Hrsg), SNG², § 10, Rz 47.

zu einem **Bewegungsprofil**, Daten über Gespräche oder Nachrichten nicht zu einem **Kommunikationsprofil** verknüpft werden,[156] dasselbe gilt für darauf gründende **Vorhersagen** über zukünftiges Verhalten.[157] Kraft der StPO mitgelesene Textnachrichten und abgehörte Telefongespräche dürfen nicht mithilfe von KI analysiert werden.[158] Gleichermaßen sind die Begriffe der Observation und der verdeckten Ermittlung nicht auf KI-Anwendungen zugeschnitten, diese sind von der Bedeutung jener Begriffe nach geltendem Recht nicht erfasst.[159]

Art 5 Abs 1 lit g KI-VO verbietet „das Inverkehrbringen, die Inbetriebnahme für diesen spezifischen Zweck oder die Verwendung von Systemen zur **biometrischen Kategorisierung**, mit denen natürliche Personen individuell auf der Grundlage ihrer biometrischen Daten kategorisiert werden, um ihre Rasse, ihre politischen Einstellungen, ihre Gewerkschaftszugehörigkeit, ihre religiösen oder weltanschaulichen Überzeugungen, ihr Sexualleben oder ihre sexuelle Ausrichtung zu erschließen oder abzuleiten; dieses Verbot gilt nicht für die **Kennzeichnung oder Filterung rechtmäßig erworbener biometrischer Datensätze**, wie z. B. Bilder auf der Grundlage biometrischer Daten oder die Kategorisierung biometrischer Daten im Bereich der Strafverfolgung".[160] Nichtsdestoweniger darf eine österreichische Behörde nur dann mit solchen KI-Systemen arbeiten, wenn dafür auch eine nationale Rechtsgrundlage besteht.

Schließlich ist zu bemerken, dass auch bei KI-unterstützten Ermittlungen – falls sie überhaupt vom österreichischen Gesetzgeber zugelassen werden – keinesfalls die „Vertraulichkeit und Integrität informationstechnischer Systeme" verletzt werden darf: Ein „Crawler"-Programm darf nicht ohne Weiteres dazu benutzt werden, Sperren, Zugangsbeschränkungen, Sicherungen etc. zu umgehen, außer Kraft zu setzen oder zu überwinden. In diesem Sinne hat der VfGH den „**Bundestrojaner**", welcher für die „Überwachung verschlüsselter Nachrichten" eingeführt worden war, für verfassungswidrig erklärt.[161]

Abschließend ist noch zu bemerken, dass es aktuelle Rechtsprechung des EuGH und des VwGH zum Begriff der „**automatisierten Entscheidung**" gibt. Zwar sind die Gerichtsentscheidungen zu Art 22 DSGVO ergangen, jedoch ist davon auszugehen, dass erstens die Terminologie sich mit Art 11 DSRL-PJ bzw. § 41 DSG deckt und zweitens gerade diese Begrifflichkeit in der KI-VO über-

156 Zu diesen Begriffen *Heißl* in *Heißl/Figl* (Hrsg), SNG², § 11, Rz 144.
157 Dabei ist jedenfalls auch das Verbot des „predictive policing" nach Art 5 Abs 1 lit d KI-VO zu beachten.
158 *Reindl-Krauskopf*, AnwBl 2024, 35 [39], sieht jedoch die Auswertung rechtmäßig sichergestellter Beweismittel mittels KI als zulässig an.
159 Dazu genauer *Heißl*, Überwachungen und Ermittlungen im Internet (2017), 56 ff, 61 ff, 68 ff, 85 ff, mwN.
160 Siehe auch Art 3 Z 40 KI-VO.
161 Vgl. VfGH 11.12.2019, G 72/2019 ua, VfSlg 20.356.

nommen oder konkretisiert wird.[162] So verbieten Art 5 Abs 3 und Art 26 Abs 10 KI-VO Entscheidungen, welche ausschließlich auf Systemen biometrischer Fernidentifizierung (ob nachträglich oder in Echtzeit) beruhen und eine nachteilige Rechtsfolge für eine Person nach sich ziehen. Wiederum verpflichtet Art 86 Abs 1 den Betreiber dazu, einer betroffenen Person eine Entscheidung im Einzelfall zu erläutern, wenn diese „auf der Grundlage der Ausgabe" eines Hochrisiko-KI-Systems[163] getroffen wurde und „rechtliche Auswirkungen" auf die Person hat oder sie „in ähnlicher Art erheblich auf eine Weise beeinträchtigt, die ihrer Ansicht nach ihre Gesundheit, ihre Sicherheit oder ihre Grundrechte beeinträchtigt". In größter kontextueller Nähe zum Problem künstlicher Intelligenz normieren Art 22 DSGVO bzw. Art 11 DSRL-PJ grundsätzlich ein Verbot der „ausschließlich auf einer automatischen Verarbeitung beruhenden Entscheidung […], die eine nachteilige Rechtsfolge für die betroffene Person hat oder sie erheblich beeinträchtigt".[164]

Nun hat der EuGH in der Sache „SCHUFA" ausgesprochen, dass eine solche Entscheidung auch dann vorliegt, wenn zuerst automatisch ein Wahrscheinlichkeitswert berechnet und dann in weiterer Folge einer Entscheidung durch einen Menschen zugrunde gelegt wird.[165] Der VwGH präzisierte im Verfahren über den „AMS-Algorithmus" diesen Gedanken dahingehend, dass er auch dann zutrifft, wenn „durch Handlungsanleitungen und Schulungen sichergestellt" wurde, dass die letztlich entscheidende Person „das Ergebnis des Algorithmus nicht unhinterfragt" übernimmt.[166] Es liegt freilich nahe, dass diese Rechtsprechung nicht nur für Wahrscheinlichkeitswerte, sondern auch für viele andere Ausgaben automatischer Systeme relevant ist, erst recht für solche von höherer Komplexität, wie beispielsweise im Falle von KI-Systemen zur Bilderkennung. Für die Themenbereiche der KI-VO bedeutet das, dass Systeme biometrischer Fernidentifizierung nicht nur einer besonders strengen und feinkörnigen menschlichen Aufsicht gemäß Art 14 unterworfen sein müssen, sondern auch in ihren Ausgaben – sofern diese nicht ohnehin zu verwerfen sind – durch weiteres, unabhängig von KI ermitteltes Material unterstützt bzw. bestätigt werden müssen. Außerdem wird bei der Beurteilung, ob eine Entscheidung gemäß Art 86 einer betroffenen Person zu erläutern sei, grundsätzlich kein enger Maßstab angelegt werden dürfen.[167]

162 Dazu *Wendehorst* in *Martini/Wendehorst*, KI-VO, Art 5, Rz 186.
163 Erfasst sind KI-Systeme nach Anhang III KI-VO mit Ausnahme des Bereichs „kritische Infrastruktur".
164 Diese Vorschriften werden in vielen Fällen gerade auf Sachverhalte anwendbar sein, in denen KI-Systeme mit personenbezogenen Daten arbeiten, zusätzlich zur KI-VO.
165 EuGH 07.12.2023, C-634/21 (*SCHUFA Holding [Scoring]*). Im konkreten Fall handelte es sich darum, dass der Wahrscheinlichkeitswert einem „Dritten" übermittelt wurde. Der VwGH übertrug diese Rechtsprechung jedoch auf eine Konstellation innerhalb einer und derselben Behörde.
166 VwGH 21.12.2023, Ra 2021/04/0010, Rz 80.
167 Diese Frage lässt sich erst mit Blick auf das einschlägige Datenschutzrecht näher untersuchen, dazu unten.

4.1. Rasterfahndung

In diesem Abschnitt geht es um die Frage, welchen Anwendungsbereich die Vorschriften der KI-VO über Hochrisiko-KI-Systeme im Bereich Strafverfolgung auf den „automationsunterstützten Datenabgleich“ nach §§ 141 ff StPO haben. Es handelt sich um ein Thema, welches in der Literatur in Ansätzen schon Beachtung gefunden hat.

§ 141 Abs 1 StPO definiert den besagten Datenabgleich, welcher kurz als „Rasterfahndung“ bezeichnet wird, folgendermaßen:

> „Im Sinne dieses Gesetzes ist „Datenabgleich“ der automationsunterstützte Vergleich von Daten (§ 36 Abs. 2 Z 1 DSG) einer Datenverarbeitung, die bestimmte, den mutmaßlichen Täter kennzeichnende oder ausschließende Merkmale enthalten, mit Daten einer anderen Datenverarbeitung, die solche Merkmale enthalten, um Personen festzustellen, die auf Grund dieser Merkmale als Verdächtige in Betracht kommen.“

Eine solche Ermittlungsmaßnahme ist gemäß Abs 2 erlaubt, wenn die Aufklärung einer mit mehr als dreijähriger Freiheitsstrafe bedrohten strafbaren Handlung „ansonsten wesentlich erschwert wäre und **nur solche Daten einbezogen** werden, die Gerichte, Staatsanwaltschaften und Sicherheitsbehörden für Zwecke eines bereits anhängigen Strafverfahrens oder sonst auf Grund bestehender Bundes- oder Landesgesetze ermittelt oder verarbeitet haben.“ Über diese Datenquellen hinaus können nach Abs 3 jedoch auch weitere Daten in den Abgleich einbezogen werden, wenn die gegenständliche Straftat mit mehr als zehn Jahren Freiheitsstrafe bedroht ist oder wegen § 278a („Kriminelle Organisation“) oder § 278b StGB („Terroristische Vereinigung“) ermittelt wird. Diesfalls können auch solchen Daten einbezogen werden, die von anderen Behörden im Wege der Amtshilfe zu übermitteln sind, ferner „Daten über Personen […], die von einem bestimmten Unternehmen bestimmte Waren oder Dienstleistungen bezogen haben oder die Mitglieder von **Personenvereinigungen** des Privatrechts oder von juristischen Personen des Privatrechts oder des öffentlichen Rechts sind.“ Abs 4 normiert ein **Verbot der Einbeziehung sensibler Daten** unter einem – eng auszulegenden – Vorbehalt:

> „Dies gilt nicht für Daten über die Staatsangehörigkeit, Daten zur tatbildmäßigen Bezeichnung einer Tätergruppe sowie für Daten, die Staatsanwaltschaften oder Sicherheitsbehörden durch erkennungsdienstliche Maßnahmen, durch Durchsuchung einer Person, durch körperliche Untersuchung oder durch molekulargenetische Analyse rechtmäßig ermittelt haben, sofern diese Daten ausschließlich für einen Datenabgleich nach Abs. 2 verwendet werden. Da-

ten von Personenvereinigungen, deren Zweck in unmittelbarem Zusammenhang mit einem der besonders geschützten Merkmale steht, dürfen in einen Datenabgleich in keinem Fall einbezogen werden.

Das bedeutet, dass zum Beispiel Daten über die ethnische Herkunft, Religionszugehörigkeit, politische oder weltanschauliche Überzeugungen von Personen nicht im Rahmen der Rasterfahndung verarbeitet werden dürfen; dasselbe gilt für Daten über einen damit verbundenen sozialen Zusammenhang, beispielsweise ein Gebetshaus oder ein politisches Forum. Dies gilt auch dann, wenn die Ziele, Motive und Vorgehensweisen einer mutmaßlichen „terroristischen Vereinigung" in untrennbarem Zusammenhang mit religiösem oder politischem Extremismus stehen. Der Schutz sensibler personenbezogener Daten bedeutet gerade, dass die betroffenen Personen nicht aufgrund der betreffenden Eigenschaften (wie zum Beispiel politischer oder religiöser Überzeugungen als solcher) erhöhter Aufmerksamkeit durch die Strafverfolgungsbehörden ausgesetzt sein sollen. Das österreichische Strafrecht kennt keine Tatbestände, deren Merkmale religiöse, politische oder weltanschauliche Überzeugungen wären.[168] Das gilt auch für die „terroristischen" Tatbilder nach §§ 278b ff und die „staatsfeindlichen Verbindungen" gemäß § 246 StGB, welche sich vielmehr durch bestimmte aggressive Bestrebungen auszeichnen, beispielsweise verfassungsmäßige Institutionen zu erschüttern oder die Bevölkerung einzuschüchtern.[169] Solche Bestrebungen sind vom Begriff der „sensiblen Daten" keineswegs erfasst und dürfen somit in einen Datenabgleich nach § 141 StPO einbezogen werden.

Eine Rasterfahndung wird gemäß § 142 StPO von der Staatsanwaltschaft angeordnet und darf ausschließlich aufgrund **gerichtlicher Bewilligung** durchgeführt werden.

Außerhalb der Strafprozessordnung ist kein automationsunterstützter Datenabgleich vorgesehen. Das bedeutet, dass er im Kontext der Sicherheitspolizei gemäß § 53 Abs 2 SPG und § 10 Abs 2 SNG nicht zulässig ist. Letztere Vorschrift gilt aber nicht für Behörden des Staatsschutzes, wenn sie gemäß § 1 Abs 4 SNG im Zusammenhang mit verfassungsgefährdenden Angriffen **kriminalpolizeiliche Aufgaben** wahrnehmen, da ihre Tätigkeit insofern ausschließlich nach der StPO zu beurteilen ist.[170]

168 Vgl. RIS-Justiz RS0088761.

169 Entsprechende Überlegungen gelten auch für die „nationalsozialistische Vereinigung" nach § 3a VerbotsG und sogar für den jüngeren Tatbestand der „religiös motivierten extremistischen Verbindung" nach § 247b StGB.

170 Dazu näher *Heißl* in *Heißl/Figl* (Hrsg), SNG², § 1, Rz 23 ff. Anzuknüpfen ist an den Anfangsverdacht nach § 1 Abs 2 und 3 StPO. Vgl. auch *Figl/Müller*, Verfassungsschutz neu: Ist der Nachrichtendienst (weiterhin) Kriminalpolizei?, JSt 2022, 135.

Sind also die Vorschriften der KI-VO über Hochrisikosysteme auf die Rasterfahndung anwendbar? Auf den ersten Blick scheint der Beitrag von *Reindl-Krauskopf* eine abschlägige Antwort auf diese Frage nahezulegen:

> „Gemessen an § 41 DSG könnte es sich hier um eine automatisierte Entscheidung im Einzelfall handeln, auch wenn das Ergebnis des Abgleichs lediglich ein weiterer Ermittlungsansatz und noch kein Beweis ist, dass eine bestimmte Person tatsächlich der Täter ist. […] Allerdings verlangt § 41 DSG, dass ein voll automatisiertes Profiling gesetzlich ausdrücklich vorgesehen werden muss. Da sich §§ 141–143 StPO zur Automatisierung nicht explizit äußern, bieten sie gerade keine ausdrückliche gesetzliche Grundlage. **Eine vollautomatisierte Rasterfahndung scheidet derzeit daher aus**."[171]

Das bedeutet, dass der automationsunterstützte Datenabgleich nicht mit neuartigen KI-Technologien durchgeführt werden darf, das Instrumentarium der Strafverfolgungsbehörden kann insofern nicht erweitert werden. Es stellt sich allerdings die Frage, ob nicht die Rasterfahndung in der Form, wie sie bereits nach bestehender Rechtslage vorgesehen ist, dem weiten Begriff des KI-Systems nach der KI-VO unterliege und ein Hochrisiko-KI-System darstelle. Ob die Rasterfahndung *de lege lata* einen KI-Einsatz darstelle, hängt von den konkreten Möglichkeiten der eingesetzten Technologien und vor allem von einer näheren Klärung des KI-Begriffs durch die Europäische Kommission oder den EuGH ab und kann hier nicht beantwortet werden – eine bejahende Antwort ist bis dahin jedenfalls nicht auszuschließen. Ist das der Fall, dann gilt die Rasterfahndung zweifelsohne als ein Hochrisiko-KI-System im Bereich „Strafverfolgung", weil sie zur Erstellung von Profilen über natürliche Personen nach Anhang III Z 6 lit e KI-VO in Verbindung mit Art 3 Z 4 DSRL-PJ dient. Sie gilt nämlich als „**Profiling**" bzw. „**automatisierte Entscheidungsfindung im Einzelfall**" im Sinne des Datenschutzrechts,[172] sodass sie nicht zu den Ausnahmen nach Art 6 Abs 3 KI-VO gehört.

Weiters stellen sich im Bereich der österreichischen StPO Abgrenzungsfragen, die für die KI-VO nicht oder nur weniger relevant sein dürften. Der „automationsunterstützte Datenabgleich" wird in der Literatur als die Bildung einer Schnittmenge aus zuvor angelegten Teilmengen von Daten definiert.[173] Davon wird die Verwendung von **Suchmaschinen** abgegrenzt, welche nach wohl überwie-

171 *Reindl-Krauskopf*, Grundrechtliche und strafprozessrechtliche Grenzen: Welche rote Linien zieht das Recht?, AnwBl 2024, 35 [39 f]. Hervorhebung M.C. Vgl. auch *Gölly/Kaiser*, Strafrecht grenzenlos? Von Grenzen und ihren Überschreitungen, JSt 2024, 298 [299], welche dieser Ansicht folgen.

172 Vgl. *Thiele/Wagner*, Praxiskommentar zum Datenschutzgesetz (DSG)² § 41 (Stand 1.2.2022, rdb.at), Rz 15; *Lachmayer* in *Knyrim,* DatKomm § 41 DSG (Stand 7.5.2020, rdb.at), Rz 14, sowie *Reindl-Krauskopf*, AnwBl 2024, 35 [39 f]; *Salimi*, Gefährliche Gruppierungen. Gruppierungen als Sicherheitsbedrohung. Grenzen und Alternativen im Kampf gegen Terror und organisierte Kriminalität durch Strukturermittlungen und Strafe (2022), Rz 158.

173 „Entscheidendes Gewicht für die Qualifizierung einer Maßnahme als ADA liegt auf dem Vergleich von Teilmengen an Daten, die aus den einzelnen Anwendungen gewonnen werden. Letztendlich erfolgt eine Schnittmengenbildung.

gender Ansicht nicht als Rasterfahndung zählt.[174] Gleichermaßen gehört die automationsunterstützte Suche innerhalb einer und derselben Datenbank nicht zu § 141 StPO.[175] Wird dabei jedoch KI eingesetzt, dann handelt es sich – unabhängig von der Qualifikation nach dem nationalen Recht – um ein Hochrisikosystem im Sinne der KI-VO, sofern dabei im Rahmen der „Strafverfolgung" gemäß Art 3 Z 46 Profile über natürliche Personen erstellt oder bearbeitet werden. Wird nur eine „auflistende und analysierende Darstellung des in einem Strafverfahren gewonnenen kriminalpolizeilichen Ermittlungsmaterials unter Verwendung eines Software-Programms" erstellt, die ebenfalls von der Rasterfahndung unterschieden wird,[176] könnte es sich um einen Fall für die Ausnahme nach Art 6 Abs 3 KI-VO handeln. Dies setzt voraus, dass eine solche Anwendung nur eine „eng gefasste Verfahrensaufgabe" durchführt – beispielsweise wenn sie „unstrukturierte Daten in strukturierte Daten umwandelt"[177] –, das „Ergebnis einer zuvor abgeschlossenen menschlichen Tätigkeit" verbessert oder eine „vorbereitende Aufgabe für eine Bewertung" wahrnimmt.[178] Werden dabei jedoch personenbezogene Daten im Rahmen einer konkreten sicherheits- oder kriminalpolizeilichen Ermittlung verarbeitet, ist im Zweifel davon auszugehen, dass diese Tätigkeit Teil der Erstellung eines Profils einer Person ist, sodass die Ausnahme des Art 6 Abs 3 nicht zum Tragen kommt.

4.2. biometrische Echtzeit-Fernidentifizierungssysteme

Art 5 Abs 1 lit h KI-VO **verbietet** grundsätzlich die „Verwendung biometrischer Echtzeit-Fernidentifizierungssysteme in **öffentlich zugänglichen Räumen** zu Strafverfolgungszwecken".[179] Jedoch werden davon im selben Atemzug weitreichende Ausnahmen gemacht, und die folgenden sechs Ab-

Die Teilmengen werden nach bestimmten Suchkriterien angelegt. Darin liegt noch kein ADA, denn innerhalb einer Datenverarbeitung werden lediglich Daten nach bestimmten Suchkriterien ausgefiltert. Erst der elektronische Abgleich der Teilmengen, also das Überprüfen, ob in den Teilmengen Überschneidungen zu finden sind, macht die besondere Qualität des ADA aus und zeichnet ihn dadurch gegenüber herkömmlichen Suchvorgängen aus" (*Reindl-Krauskopf* in *Fuchs/Ratz*, WK StPO § 141 (Stand 1.11.2019, rdb.at), Rz 9/1. Dort mit Hervorhebungen).

174 Vgl. ebd., Rz 13/1; *Venier* in *Bertel/Venier* (Hrsg), StPO: Kommentar² (2022) zu § 141, Rz 1; *Salimi*, Gefährliche Gruppierungen, Rz 158 ff; anderer Ansicht *Heißl* in *Heißl/Figl* (Hrsg), SNG², § 10, Rz 15 ff; *ders.*, Darf die Polizei googeln? Der Gebrauch des World Wide Web in sicherheitspolizeilichen Ermittlungen, ZfV 2016, 158 [160 f]; *ders.*, Überwachungen und Ermittlungen im Internet (2017), 90 ff.

175 Vgl. *Reindl-Krauskopf* in *Fuchs/Ratz*, WK StPO § 141 (Stand 1.11.2019, rdb.at), Rz 13; *Ohrnhofer* in *Schmölzer/Mühlbacher* (Hrsg), StPO Kommentar, Band 1 Ermittlungsverfahren² (2021) § 141 StPO, Rz 3; *Rohregger* in *Birklbauer/Haumer/Nimmervoll/Wess* (Hrsg), StPO – Linzer Kommentar zur Strafprozessordnung (2020) zu § 141 StPO, Rz 4.

176 OLG Wien 21 Bs 21/00, JBl 2001, 257.

177 ErwGr 53 KI-VO.

178 Vgl. auch ErwGr 53 KI-VO.

179 ErwGr 32 KI-VO: „Die Verwendung von KI-Systemen zur biometrischen Echtzeit-Fernidentifizierung natürlicher Personen in öffentlich zugänglichen Räumen zu Strafverfolgungszwecken greift besonders in die Rechte und Freiheiten der betroffenen Personen ein, da sie die Privatsphäre eines großen Teils der Bevölkerung beeinträchtigt, ein **Gefühl der ständigen Überwachung** weckt und indirekt von der Ausübung der **Versammlungsfreiheit** und anderer Grundrechte abhalten kann."

sätze widmen sich der Konkretisierung dieser Erlaubnis.[180] Demnach ist die Verwendung eines solchen Systems zulässig, „wenn und insoweit dies im Hinblick auf eines der folgenden Ziele unbedingt erforderlich ist:

> i) gezielte **Suche nach bestimmten Opfern** von Entführung, Menschenhandel oder sexueller Ausbeutung sowie die Suche nach **vermissten Personen**;
> ii) Abwenden einer konkreten, erheblichen und unmittelbaren **Gefahr für das Leben oder die körperliche Unversehrtheit** natürlicher Personen oder einer tatsächlichen und bestehenden oder tatsächlichen und vorhersehbaren Gefahr eines **Terroranschlags**;[181]
> iii) Aufspüren oder Identifizieren einer Person, die der Begehung einer **Straftat verdächtigt** wird, zum Zwecke der Durchführung von strafrechtlichen Ermittlungen oder von Strafverfahren oder der Vollstreckung einer Strafe für die in **Anhang II** aufgeführten Straftaten, die in dem betreffenden Mitgliedstaat nach dessen Recht mit einer **Freiheitsstrafe** oder einer freiheitsentziehenden Maßregel der Sicherung im Höchstmaß von **mindestens vier Jahren** bedroht ist."

Nach Anhang II handelt es sich dabei um folgende Straftaten: Terrorismus, Menschenhandel, sexuelle Ausbeutung von Kindern und Kinderpornografie, illegaler Handel mit Drogen oder psychotropen Stoffen, illegaler Handel mit Waffen, Munition oder Sprengstoffen, Mord, schwere Körperverletzung, illegaler Handel mit menschlichen Organen oder menschlichem Gewebe, illegaler Handel mit nuklearen oder radioaktiven Substanzen, Entführung, Freiheitsberaubung oder Geiselnahme, Verbrechen, die in die Zuständigkeit des Internationalen Strafgerichtshofs fallen, Flugzeug- oder Schiffsentführung, Vergewaltigung, Umweltkriminalität, organisierter oder bewaffneter Raub, Sabotage, Beteiligung an einer kriminellen Vereinigung, die an einer oder mehreren der oben genannten

180 Daran kann freilich Kritik geübt werden, insbesondere weil die Vorschrift wie ein inkonsequenter Kompromiss wirkt und die Übersichtlichkeit der Vorschriften über „*verbotene* Praktiken" durch die weitreichende Erlaubnis samt ausführlichen Verfahrensvorschriften gestört wird. Vgl. auch *Wendehorst* in *Martini/Wendehorst*, KI-VO (2024), Art 5, Rz 136 ff.

181 Zur Präzisierung ErwGr 33 KI-VO: „Eine unmittelbare Gefahr für das Leben oder die körperliche Unversehrtheit natürlicher Personen kann auch durch die schwerwiegende **Störung einer kritischen Infrastruktur** […] entstehen, wenn die Störung oder Zerstörung einer solchen kritischen Infrastruktur zu einer unmittelbaren Gefahr für das Leben oder die körperliche Unversehrtheit einer Person führen würde, auch durch die schwerwiegende Beeinträchtigung der Bereitstellung der **Grundversorgung** für die Bevölkerung oder der Wahrnehmung der Kernfunktion des Staates. Darüber hinaus sollte diese Verordnung die Fähigkeit der Strafverfolgungs-, Grenzschutz-, Einwanderungs- oder Asylbehörden erhalten, gemäß den im Unionsrecht und im nationalen Recht für diesen Zweck festgelegten Bedingungen die Identität der betreffenden Person in ihrer Anwesenheit festzustellen. Insbesondere sollten Strafverfolgungs-, Grenzschutz-, Einwanderungs- oder Asylbehörden gemäß dem Unionsrecht oder dem nationalen Recht Informationssysteme verwenden können, um eine Person zu identifizieren, die **während einer Identitätsfeststellung** entweder verweigert, identifiziert zu werden, oder nicht in der Lage ist, ihre Identität anzugeben oder zu belegen, wobei gemäß dieser Verordnung keine vorherige Genehmigung erlangt werden muss. Dabei könnte es sich beispielsweise um eine Person handeln, die in eine Straftat verwickelt ist und nicht gewillt oder aufgrund eines Unfalls oder des Gesundheitszustands nicht in der Lage ist, den Strafverfolgungsbehörden ihre Identität offenzulegen."

Straftaten beteiligt ist. Diese Begriffe sind nicht nach nationalem Strafrecht, sondern autonom nach dem Unionsrecht auszulegen.[182]

Das „biometrische Fernidentifizierungssystem“ überhaupt – nachträglich oder in Echtzeit – wird in Art 3 Z 41 KI-VO definiert als „KI-System, das dem Zweck dient, natürliche Personen **ohne ihre aktive Einbeziehung** und in der Regel aus der Ferne durch **Abgleich** der biometrischen Daten einer Person mit den in einer **Referenzdatenbank** gespeicherten biometrischen Daten zu identifizieren“.[183] Das „biometrische **Echtzeit**-Fernidentifizierungssystem“ ist laut Z 42 „ein biometrisches Fernidentifizierungssystem, bei dem die Erfassung biometrischer Daten, der Abgleich und die Identifizierung **ohne erhebliche Verzögerung** erfolgen, und das zur Vermeidung einer Umgehung der Vorschriften nicht nur die sofortige Identifizierung, sondern auch eine Identifizierung mit **begrenzten kurzen Verzögerungen** umfasst“. Erfolgt die Identifikation mit einer längeren Verzögerung, handelt es sich gemäß Z 43 bloß um ein „System zur nachträglichen biometrischen Fernidentifizierung“. Wann genau sie „ohne erhebliche Verzögerung“ bzw. nur „mit begrenzten kurzen Verzögerungen“ stattfindet,[184] wird sich nach dem jeweiligen Verwendungskontext richten und sollte im Verlauf der Praxis in Form von Urteilen, Erkenntnissen etc. oder bestenfalls von Leitlinien präzisiert werden. Die Identifizierung in Echtzeit stellt für die betroffene Person ein besonders großes Risiko und einen besonders schwerwiegenden Grundrechtseingriff dar, weil sie dann sofort beispielsweise festgenommen werden kann, während sie bei einer bloß nachträglichen Identifizierung mitunter die Möglichkeit haben wird, sich den Behörden zu entziehen.[185]

182 Siehe ErwGr 33 KI-VO; Rahmenbeschluss 2002/584/JI des Rates vom 13. Juni 2002 über den Europäischen Haftbefehl und die Übergabeverfahren zwischen den Mitgliedstaaten (ABl. L 190 vom 18.7.2002, S. 1).

183 Näher dazu ErwGr 15 KI-VO: „Der Begriff 'biometrische Identifizierung' sollte gemäß dieser Verordnung als automatische Erkennung physischer, physiologischer und verhaltensbezogener menschlicher Merkmale wie Gesicht, Augenbewegungen, Körperform, Stimme, Prosodie, Gang, Haltung, Herzfrequenz, Blutdruck, Geruch, charakteristischer Tastenanschlag zum Zweck der Überprüfung der Identität einer Person durch Abgleich der biometrischen Daten der entsprechenden Person mit den in einer Datenbank gespeicherten biometrischen Daten definiert werden, unabhängig davon, ob die Einzelperson ihre Zustimmung dazu gegeben hat oder nicht. **Dies umfasst keine** KI-Systeme, die bestimmungsgemäß für die biometrische Verifizierung, wozu die Authentifizierung gehört, verwendet werden sollen, deren einziger Zweck darin besteht, zu bestätigen, dass eine bestimmte natürliche Person die Person ist, für die sie sich ausgibt, sowie zur Bestätigung der Identität einer natürlichen Person zu dem alleinigen Zweck Zugang zu einem Dienst zu erhalten, ein Gerät zu entriegeln oder Sicherheitszugang zu Räumlichkeiten zu erhalten.“

184 Auch ErwGr 17 lässt eine höhere begriffliche Präzision vermissen: „'Echtzeit-Systeme' umfassen die Verwendung von **„Live-Material“ oder „Near-live-Material“** wie etwa Videoaufnahmen, die von einer Kamera oder einem anderen Gerät mit ähnlicher Funktion erzeugt werden. Bei Systemen zur nachträglichen Identifizierung hingegen wurden die biometrischen Daten schon zuvor erfasst und der Abgleich und die Identifizierung erfolgen erst mit erheblicher Verzögerung. Dabei handelt es sich um Material wie etwa Bild- oder Videoaufnahmen, die von Video-Überwachungssystemen oder privaten Geräten vor der Anwendung des Systems auf die betroffenen natürlichen Personen erzeugt wurden.“

185 *Wendehorst* in *Martini/Wendehorst*, KI-VO, Art 5, Rz 144, bemerkt kritisch, dass die nachträgliche Identifizierung einerseits aufgrund der Notwendigkeit, Aufnahmen zu speichern, ein höheres Datenschutzrisiko birgt und andererseits aufgrund des zeitlichen Intervalls zwischen Aufzeichnung und Identifikation, daran anschließend behördlichem Eingreifen, ineffektiv und daher unverhältnismäßig sein könnte.

Wird die betroffene Person etwa aufgefordert, Fingerabdrücke zu geben oder sich scannen zu lassen, liegt keine biometrische Echtzeit-FI vor, weil diese „ohne ihre aktive Einbeziehung“ stattfindet.[186] Das Tatbestandsmerkmal „öffentlich zugänglicher Raum“ in Abs 5 Abs 1 lit h KI-VO bestimmt Art 3 Z 44 als „einen einer unbestimmten Anzahl natürlicher Personen zugänglichen physischen Ort in privatem oder öffentlichem Eigentum, unabhängig davon, ob bestimmte Bedingungen für den Zugang gelten, und unabhängig von möglichen Kapazitätsbeschränkungen“. Dieser Begriff ist nach ErwGr 19 weit auszulegen.[187]

Art 5 Abs 2 KI-VO enthält eine zentrale Einschränkung der Befugnis zur biometrischen Echtzeit-FI: Sie darf ausschließlich „zur **Bestätigung der Identität der speziell betroffenen Person** erfolgen“. Eine ungezielte oder anlasslose Überwachung und Fernerkennung von Personen in Echtzeit ist somit prinzipiell ausgeschlossen und daher eine verbotene Praktik im Sinne des Art 5 Abs 1 – dies gilt auch dann, wenn der überwachte Ort oder Kontext als besonders gefährlich gilt oder es tatsächlich ist.[188] Das bedeutet, dass auf jeden Fall nur nach einer endlichen Anzahl an bestimmten Personen auf die beschriebene Weise gesucht werden darf, dass also die relevanten Datenmengen, mit denen abgeglichen wird, im voraus abzugrenzen und nach bestimmten Personen zu definieren sind.[189] Diese Personen können ausschließlich die in Art 5 Abs 1 lit h aufgezählten sein, mithin Opfer bestimmter Straftaten, Vermisste, Verdächtige von Verbrechen nach Anhang II und Menschen, von denen eine konkrete, erhebliche und unmittelbare Gefahr für das Leben bzw. die körperliche Unversehrtheit oder eines Terroranschlags ausgeht.[190] Es erscheint zudem geboten, diese Auswahl der Datensätze genau und differenziert schriftlich zu **begründen**. In diesem Zusammenhang liegt es nahe, die **automatisch erzeugten** (bzw. zu erzeugenden) **Protokolle** nach Art 12 Abs 3 KI-VO, welche mindestens die „Aufzeichnung jedes Zeitraums der Verwendung des Systems“, die Referenzdatenbank, die Eingabedaten und die Identität der gemäß Art 14 Abs 5 mit der menschlichen Aufsicht Befassten

186 Vgl. *dies.*, ebd., Art 3, Rz 296.

187 Dieser nennt „Bereiche wie etwa für Gewerbe, etwa Geschäfte, Restaurants, Cafés, für Dienstleistungen, etwa Banken, berufliche Tätigkeiten, Gastgewerbe, für Sport, etwa Schwimmbäder, Fitnessstudios, Stadien, für Verkehr, etwa Bus- und U-Bahn-Haltestellen, Bahnhöfe, Flughäfen, Transportmittel, für Unterhaltung, etwa Kinos, Theater, Museen, Konzert- und Konferenzsäle oder für Freizeit oder Sonstiges, etwa öffentliche Straßen und Plätze, Parks, Wälder, Spielplätze.“ Vgl. auch allgemein zur Thematik der Öffentlichkeit: *Zanol*, Über den Schutzbereich öffentlich verfügbarer Daten im Anwendungsbereich der europäischen Datenschutz-Grundverordnung (DSGVO) (Dissertation, Uni Wien, unveröffentlicht).

188 Vgl. *Wendehorst* in *Martini/Wendehorst*, KI-VO, Art 5, Rz 155, 165.

189 Ähnlich *Wendehorst* in *Martini/Wendehorst*, KI-VO, Art 5, Rz 163: „Dies ist deswegen so wichtig zu betonen, weil ein verbreitetes Fehlverständnis offen davon ausgeht, dass bei der biometrischen Echtzeit-Fernidentifizierung in öffentlich zugänglichen Räumen jeder einzelne Passant identifiziert wird. […] Das bedeutet, dass **in der Referenzdatenbank nur jene Profile freigeschaltet werden dürfen, in Bezug auf die eine Maßnahme** [nach dieser Vorschrift] **gerechtfertigt** ist.“ (Hervorhebung nach Original)

190 Vgl. ebd., Rz 165: Nicht umfasst sind etwa Personen, die bloß als Zeugen in Frage kommen.

beinhalten, im Blick zu behalten und möglichst jeden Arbeits- bzw. Denkschritt in die automatische Dokumentation einzubeziehen.

Darüber hinaus ist die biometrische Echtzeit-FI im Sinne der KI-VO nur dann zulässig, wenn eine strenge **Verhältnismäßigkeitsprüfung** für sie spricht.[191] Demnach muss nach Art 5 Abs 2 immer vor der Verwendung des Systems geprüft werden:

> „a) die Art der **Situation**, die der möglichen Verwendung zugrunde liegt, insbesondere die Schwere, die Wahrscheinlichkeit und das Ausmaß des **Schadens**, der entstehen würde, **wenn das System nicht eingesetzt würde**;
> b) die **Folgen** der Verwendung des Systems für die **Rechte und Freiheiten** aller betroffenen Personen, insbesondere die Schwere, die Wahrscheinlichkeit und das Ausmaß solcher Folgen."

Die Folgen der KI-Verwendung zur Echtzeit-FI im Bereich Strafverfolgung werden in Schwere, Wahrscheinlichkeit und Ausmaß dadurch vermindert, dass nach dieser Vorschrift „notwendige und verhältnismäßige **Schutzvorkehrungen und Bedingungen** für die Verwendung im Einklang mit **nationalem Recht** über die Ermächtigung ihrer Verwendung" eingehalten werden, „insbesondere in Bezug auf die zeitlichen, geografischen und personenbezogenen Beschränkungen." Es besteht daher **Umsetzungsbedarf**, das heißt: Die österreichische Gesetzgebung muss, damit die biometrische Echtzeit-FI überhaupt möglich sei, eigene detaillierte Rechtsnormen erlassen. Diese müssen dafür sorgen, dass die „Rechte und Freiheiten **aller betroffenen Personen**" im Rahmen des Betriebs solcher Systeme geschützt werden. Die betreffenden Gesetze müssen vorgeben, dass die Ermächtigung zu dergleichen Ermittlungsmaßnahmen nur für genau abgegrenzte, überschaubare Zeiträume bestehen kann und sich auf genau anzugebende Orte beschränken muss.

Zum Schutz der **direkt** Betroffenen – also der jeweils gesuchten Personen – müssen jedenfalls die nötigen Datenschutz-Vorkehrungen getroffen werden: Dazu gehört insbesondere das Recht, keiner Entscheidung ausgesetzt zu werden, die eine nachteiligen Rechtsfolge „ausschließlich auf der Grundlage der Ausgabe des biometrischen Echtzeit-Fernidentifizierungssystems" nach sich zieht. Ein solches Recht ist bereits aus Art 11 DSRL-PJ – in § 41 DSG nur unzureichend umgesetzt[192] – bekannt und wird ein weiteres Mal im letzten Satz des Art 5 Abs 3 KI-VO übernommen bzw. kon-

191 Vgl. ebd., Rz 166 ff.

192 Der Text dieser Norm lässt das „Recht auf persönliches Eingreifen seitens des Verantwortlichen" bzw. „geeignete Garantien für die Rechte und Freiheiten der betroffenen Person" gemäß Art 11 Abs 1 DSRL-PJ schlichtweg aus und muss daher im Wege richtlinienkonformer Auslegung berichtigt werden. Vgl. *Lachmayer* in *Knyrim,* DatKomm § 41 DSG (Stand 7.5.2020, rdb.at), Rz 4.

kretisiert.[193] Das bedeutet, dass die – vorwiegend technisch verstandene – menschliche Aufsicht gemäß Art 14 Abs 5 hier von besonderer Notwendigkeit ist und zugleich alle Ausgaben des KI-Systems im Rahmen einer sorgfältigen Aktenführung, welche jedes Für und Wider berücksichtigt, und rechtskundigen Beurteilung kritisch überprüft werden müssen.[194] Bei den genannten direkt betroffenen Personen ist insbesondere an die Unschuldsvermutung, allgemein das Recht auf ein faires Verfahren nach Art 6 EMRK und Art 47 ff GRC, das Diskriminierungsverbot hinsichtlich ethnischer und kultureller Merkmale und nicht zuletzt an datenschutzrechtliche Erfordernisse[195] zu denken. Bei **unbeteiligten Dritten** erscheint auf ganz fundamentaler Ebene das Recht auf Privatsphäre gemäß Art 8 EMRK sowie Art 7 GRC und der Schutz vor sachlich unbegründeten Strafverfolgungsmaßnahmen essenziell.[196]

Nach Art 5 Abs 2 KI-VO darf ein solches System erst dann in Betrieb genommen werden, nachdem eine besondere **Grundrechte-Folgenabschätzung** durchgeführt wurde. Zudem muss das System in der **EU-Datenbank** nach Art 49 registriert sein.

Für die Verwendung von Systemen der biometrischen Echtzeit-FI in öffentlich zugänglichen Räumen zu Strafverfolgungszwecken normiert Art 5 Abs 3 KI-VO den „**Richtervorbehalt**“: Demnach bedarf es dafür der **vorherigen Genehmigung eines Gerichts oder einer unabhängigen Verwaltungsbehörde**.[197] Im Rahmen der StPO wird es sich dabei jedenfalls um das Landesgericht als Rechtsschutzrichter im Ermittlungsverfahren handeln müssen, im Anwendungsbereich des SPG bzw. SNG müssten jedoch noch Überlegungen nach der richtigen Behörde angestellt werden. Die Stellung und die Befugnisse des Rechtsschutzbeauftragen dürften nicht hinreichend sein, um ihn als

193 *Wendehorst* in *Martini/Wendehorst*, KI-VO, Art 5, Rz 184 ff. Auch hier hat sich der EU-Verordnungsgeber für eine überraschende Stelle entschieden und der Übersichtlichkeit dadurch Eintrag getan.

194 ErwGr 32 KI-VO: „**Technische Ungenauigkeiten** von KI-Systemen, die für die biometrische Fernidentifizierung natürlicher Personen bestimmt sind, können zu verzerrten Ergebnissen führen und eine **diskriminierende** Wirkung haben. Solche möglichen verzerrten Ergebnisse und eine solche diskriminierende Wirkung sind von besonderer Bedeutung, wenn es um das Alter, die ethnische Herkunft, die Rasse, das Geschlecht oder Behinderungen geht. Darüber hinaus bergen die Unmittelbarkeit der Auswirkungen und die **begrenzten Möglichkeiten weiterer Kontrollen oder Korrekturen** im Zusammenhang mit der Verwendung solcher in Echtzeit betriebener Systeme erhöhte Risiken für die Rechte und Freiheiten der betreffenden Personen, die im Zusammenhang mit Strafverfolgungsmaßnahmen stehen oder davon betroffen sind.“

195 Zu den datenschutzrechtlichen Aspekten der Videoüberwachung und biometrischen Erkennung im Allgemeinen siehe beispielsweise: *Kastelitz/Tschohl/Hötzendorfer*, (Datenschutz-)Rechtliche Aspekte der polizeilichen Verarbeitung von Videomassendaten, JBD 2019, 327; *Schwaiger*, Biometrische Gesichtserkennung, JBD 2016, 193; *Pollirer*, Checkliste Videoüberwachung nach der EDPB-Leitlinie 3/2019, Dako 2021, 12.

196 Vgl. zu dem Ganzen *Wendehorst* in *Martini/Wendehorst*, KI-VO, Art 5 , Rz 169 f, 172 ff.

197 „Die zuständige Justizbehörde oder eine unabhängige Verwaltungsbehörde, deren Entscheidung bindend ist, erteilt die Genehmigung nur dann, wenn sie auf der Grundlage objektiver Nachweise oder eindeutiger Hinweise, die ihr vorgelegt werden, davon überzeugt ist, dass die Verwendung des betreffenden biometrischen Echtzeit-Fernidentifizierungssystems für das Erreichen eines der in Absatz 1 Unterabsatz 1 Buchstabe h genannten Ziele — wie im Antrag angegeben — notwendig und verhältnismäßig ist und insbesondere auf das in Bezug auf den Zeitraum sowie den geografischen und persönlichen Anwendungsbereich unbedingt erforderliche Maß beschränkt bleibt.“ (Art 5 Abs 3 KI-VO)

„unabhängige Verwaltungsbehörde“ qualifizieren zu können.[198] Das Gleiche muss auch für den Rechtsschutzsenat nach § 14 Abs 3 SNG gelten, welchem der RSB und zwei Stellvertreter angehören. Zu erwägen wäre es daher – falls in Österreich der Einsatz solcher Systeme überhaupt erwünscht ist[199] –, die **Datenschutzbehörde** für die Genehmigung zuständig zu machen.[200] Ferner bestünde die Möglichkeit, die Genehmigung durch ein **Verwaltungsgericht** (VwG) vorzusehen. Zwar sind die Verwaltungsgerichte in erster Linie für Beschwerden über Entscheidungen und andere Handlungen bzw. Unterlassungen von Verwaltungsbehörden zuständig – sie haben insofern nur die Überprüfung bereits bestehender Fakten zum Gegenstand –, doch enthält Art 130 Abs 2 Z 4 B-VG auch die Möglichkeit, die VwG mit der Entscheidung über „Streitigkeiten und Anträge in sonstigen Angelegenheiten“ zu betrauen.[201] Dass ein VwG, sozusagen als Behörde erster Instanz, auf Antrag über die Genehmigung für den Einsatz einer Technologie entscheidet, wäre in Österreich zwar eine systematische Neuerung, aber ausdrücklich von der Verfassung gedeckt.[202] Einerseits würde der Instanzenzug dadurch kürzer werden, weil eine Anfechtung der Entscheidung des VwG nur noch beim VwGH unter *de lege lata* engen Bedingungen möglich wäre, andererseits entschiede sofort – einen Antrag der Polizeibehörde vorausgesetzt – eine Behörde, deren Unabhängigkeit und besonders hohe rechtliche Qualifikation verfassungsmäßig garantiert ist.[203] Es kann, muss aber nicht sein, dass die Bevölkerung betreffende Maßnahmen dann eher als legitim akzeptiert.

Art 5 Abs 3 KI-VO enthält jedoch eine weitere strenge Regel für das, was die StPO als „**Gefahr im Verzug**“ bezeichnet:

> „In **hinreichend begründeten dringenden Fällen** kann jedoch mit der Verwendung eines solchen Systems zunächst ohne Genehmigung begonnen werden, sofern eine solche Genehmigung unverzüglich, spätestens jedoch innerhalb von **24 Stunden** beantragt wird. Wird eine solche Genehmigung abgelehnt, so wird die Verwendung mit sofortiger Wirkung **eingestellt** und werden alle Daten sowie die Ergebnisse und Ausgaben dieser Verwendung unverzüglich **verworfen und gelöscht**.“

198 Vgl. VwGH 08.03.2023, Ra 2022/03/0214 (RS 4); *Lachmayer*, Rechtsstaatliche Grenzen polizeilicher Überwachungsbefugnisse, JBÖR 2020, 105 [117 f], mit Verweis auf *Klaushofer*, Strukturfragen der Rechtsschutzbeauftragten (2012), 466 ff.

199 Siehe die „**Öffnungsklausel**“ in Art 5 Abs 5 KI-VO, wonach die Vorschriften der KI-VO zu diesem Thema die Behörden nicht unmittelbar zum Einsatz der Technologie ermächtigen, sondern nur einen Rahmen für die nationale Gesetzgebung („Umsetzung“) vorgeben.

200 Vgl. *Salimi*, Gefährliche Gruppierungen, Rz 978 ff.

201 Dazu näher *Muzak,* B-VG[6] Art 130 (Stand 1.10.2020, rdb.at), Rz 13; Art 136 Abs 3b B-VG.

202 Anderer Ansicht *Salimi*, Gefährliche Gruppierungen, Rz 994, ohne auf Art 130 Abs 2 Z 4 B-VG näher einzugehen.

203 Befürwortend auch ebd., Rz 998.

ErwGr 35 spricht in diesem Zusammenhang von „Situationen, in denen es wegen der Notwendigkeit der Verwendung der betreffenden Systeme tatsächlich und **objektiv unmöglich** ist, vor dem Beginn der Verwendung des KI-Systems eine Genehmigung einzuholen. In solchen dringenden Fällen sollte die Verwendung des KI-Systems auf das **absolut notwendige Mindestmaß** beschränkt werden und angemessenen Schutzvorkehrungen und Bedingungen unterliegen, die im nationalen Recht festgelegt sind und im Zusammenhang mit jedem einzelnen dringenden Anwendungsfall von der Strafverfolgungsbehörde selbst **präzisiert** werden." Außerdem sollte die Behörde die Gründe dafür angeben, warum sie die Genehmigung nicht früher beantragen konnte.[204]

Gemäß Art 5 Abs 4 KI-VO muss die Verwendung eines solchen Systems außerdem der **Marktüberwachungsbehörde** und der **Datenschutzbehörde** mitgeteilt werden. Sensible operative Daten im Sinne des Art 3 Z 38 sind davon ausgenommen. Letztere Stellen legen der **Europäischen Kommission** Jahresberichte nach Art 5 Abs 6 vor, die Kommission informiert wiederum gemäß Abs 7 auf der Grundlage aggregierter Daten die **Öffentlichkeit**.

Eine biometrische Echtzeit-FI in öffentlich zugänglichen Räumen für die Strafverfolgung darf keineswegs unmittelbar auf die KI-VO gestützt werden. Die Behörde ist erst dann zu einem solchen Handeln imstande, wenn der Mitgliedstaat (MS) dafür selbst Gesetze samt näheren Verfahrensvorschriften erlassen hat. Diese wiederum müssen mit dem nationalen Verfassungsrecht und allen relevanten Grundrechten übereinstimmen („**doppelte Bindung**"). Art 5 Abs 5 KI-VO stellt die besprochene Technologie in diesem Sinne unter einen **Umsetzungsvorbehalt**, sie bedeutet eine **Öffnungsklausel**. Der jeweilige Mitgliedstaat muss zuerst die Frage beantworten, ob die KI-Verwendung überhaupt zulässig sei – er darf, muss sie aber nicht erlauben. Entscheidet er sich dafür, muss er durch Gesetz[205] die Reichweite der Befugnis und die „erforderlichen **detaillierten Vorschriften** für die Beantragung, Erteilung und Ausübung der […] Genehmigungen sowie für die entsprechende Beaufsichtigung und Berichterstattung" festlegen.[206] Die KI-VO gibt nur einen äußeren Rahmen vor: Der MS darf keine Vorschriften erlassen, die den Behörden weiterreichende Befugnisse einräumen, als die Verordnung es tut, er darf aber sehr wohl engere Bedingungen der Zulässigkeit sowie strengere Verfahrensvorschriften vorsehen. Er kann beispielsweise regeln, dass nur bestimmte Straftaten des Anhang II mithilfe der biometrischen Echtzeit-FI verfolgt werden. Die nationalen Vor-

204 *Wendehorst* in *Martini/Wendehorst*, Art 5, Rz 179.

205 Die KI-VO spricht hier zwar nicht von Gesetzen, weil es auch andere Rechtstraditionen gibt, in Österreich ist es aber selbstverständlich, dass dergleichen nur durch Gesetz geregelt werden kann.

206 Siehe dazu die Details im Leitlinienentwurf der Kommission: C(2025) 884 final https://ec.europa.eu/newsroom/dae/redirection/document/112367; Abruf 28.02.2025), Absatz 417.

schriften sind der Kommission bei sonstiger Unwirksamkeit[207] binnen 30 Tagen ab Erlassung mitzuteilen.

ErwGr 38 KI-VO stellt wohlgemerkt klar, dass die besprochenen Normen der Verordnung keine neue Rechtsgrundlage für die Verarbeitung personenbezogener Daten darstellen. Sämtliche Daten, welche im Rahmen des KI-Einsatzes – etwa als Trainingsdaten, Elemente einer Referenzdatenbank, Eingaben oder Ausgaben – verarbeitet werden, müssen im Sinne des Datenschutzrechts rechtmäßig erworben oder weiterverarbeitet werden. Selbst wenn das KI-System im Sinne der KI-VO rechtmäßig eingesetzt wird und deren sämtliche Vorgaben eingehalten werden, ist die Verwendung dieses Systems möglicherweise aufgrund des Datenschutzrechts trotzdem unzulässig. KI-Recht und Datenschutzrecht stellen zwei verschiedene Aspekte dar, welche unabhängig voneinander zu prüfen sind.

4.3. Systeme zur nachträglichen biometrischen Fernidentifizierung

Jedes biometrische Fernidentifizierungssystem, welches nicht nach Art 3 Z 42 KI-VO Personen in Echtzeit erkennt, ist laut Z 43 ein System zur nachträglichen biometrischen Fernidentifizierung. Es kann dabei zwischen erstmaliger Aufzeichnung oder Veröffentlichung der biometrischen Daten somit ein unbegrenzter Zeitraum liegen. Solche Systeme werden, sofern sie „im Rahmen von Ermittlungen zur gezielten Suche einer Person, die der Begehung einer Straftat verdächtigt wird oder aufgrund einer solchen verurteilt wurde“, verwendet werden, in Art 26 Abs 10 KI-VO (Überschrift „Pflichten der Betreiber von Hochrisiko-KI-Systemen“) geregelt.[208] Der Unterschied zu den Vorschriften über Systeme der Echtzeit-FI besteht zum Beispiel in einigen verfahrensrechtlichen Lockerungen. Doch bereits der Anwendungsbereich ist weiter gefasst, da im Gegensatz zu Art 5 Abs 2 in Art 26 Abs 10 keine Einschränkung auf einen Systemeinsatz in „öffentlich zugänglichen Räumen“ getroffen wird. Das heißt, dass nicht nur Überwachungen wie solche durch **Videokameras oder Sensoren**, sondern auch beispielsweise **Technologien zur Identifikation von Personen in Internetvideos, Fotos oder Audioaufnahmen** Gegenstand dieser Vorschrift sind. Es ist gleichermaßen denkbar, dass mit einschlägigen KI-Systemen Personen aufgrund von bloßen Texten identifiziert werden, falls diese auf inhaltlicher Ebene biometrische Daten enthalten oder bereits in formaler Hinsicht – sei es stilistisch, graphologisch oder was das Textbild betrifft – kraft Biometrie Auf-

207 Vgl. ebd., Absatz 415, mit Verweis auf die Rechtsprechung des EuGH.

208 Diese Regelungen sind an der vom Verordnungsgeber gewählten Stelle fehl am Platz, weil unter den allgemeinen Betreiberpflichten überraschender Weise eine sehr spezifische sowie heikle Materie ausführlich geregelt wird. Es wäre weitaus übersichtlicher, deutlicher und nachvollziehbarer gewesen, wenn man die Echtzeit- sowie die nachträgliche FI jeweils in eigenen Artikeln in textueller Nähe zueinander geregelt hätte. Hieran zeigt sich eindrücklich, dass legislative Kompromisse einen Rechtstext in formaler Hinsicht empfindlich beeinträchtigen können.

schluss über die Identität einer Person geben könnten. Die KI-VO geht nämlich von einem weiten Begriff biometrischer Daten aus, indem sie diese in Art 3 Z 34 als „mit speziellen technischen Verfahren gewonnene personenbezogene Daten zu den **physischen, physiologischen oder verhaltenstypischen Merkmalen** einer natürlichen Person“ definiert.[209]

Nach Art 26 Abs 10 KI-VO „beantragt der Betreiber eines Hochrisiko-KI-Systems zur nachträglichen biometrischen Fernfernidentifizierung im Rahmen von Ermittlungen zur gezielten Suche einer Person, die der Begehung einer Straftat verdächtigt wird oder aufgrund einer solchen verurteilt wurde, vorab oder unverzüglich, spätestens jedoch binnen **48 Stunden** bei einer **Justizbehörde** oder einer **Verwaltungsbehörde**, deren Entscheidung bindend ist und einer justiziellen Überprüfung unterliegt, die **Genehmigung** für die Nutzung dieses Systems, **es sei denn**, es wird zur **erstmaligen Identifizierung eines potenziellen Verdächtigen auf der Grundlage objektiver und nachprüfbarer Tatsachen, die in unmittelbarem Zusammenhang mit der Straftat stehen**, verwendet. Jede Verwendung ist auf das für die Ermittlung einer bestimmten Straftat **unbedingt erforderliche Maß** zu beschränken.“ Wird die Genehmigung versagt, ist das System einzustellen und personenbezogene Daten sind zu löschen, „die im Zusammenhang mit der Verwendung des Hochrisiko-KI-Systems stehen“.

Die wichtigsten Unterschiede zu den Regeln über die biometrische Echtzeit-FI bestehen somit in folgenden Punkten: Erstens darf die Technologie unterschiedslos, also unabhängig von den Fällen des Art 5 Abs 1 lit h, für Ermittlungen im Bereich „Strafverfolgung“ verwendet werden; zweitens darf das KI-System vor der Genehmigung nicht nur in „hinreichend begründeten dringenden Fällen“ in Betrieb genommen werden, und der diesbezügliche Antrag ist nicht innerhalb von 24, sondern von 48 Stunden zu stellen; drittens braucht im Falle der Genehmigung durch eine Verwaltungsbehörde diese nicht „unabhängig“ zu sein, sondern ihre Entscheidung muss lediglich einer „justiziellen Überprüfung“ zugänglich sein. Fernerhin fehlen Bestimmungen über eine Eintragung in die EU-Datenbank und über die GRFA. Zu bemerken ist auch, dass die Verordnung an dieser Stelle keine Umsetzung durch Mitgliedstaaten thematisiert – jedoch wird den MS freigestellt, dafür **strengere Vorschriften** zu erlassen, die inhaltlicher oder prozessualer Natur sein können. Aus dem Text lässt sich ferner leicht erschließen, dass der MS jedenfalls die Behörde und das Verfahren für die Genehmigung selbst festlegen muss. Außerdem spricht Anhang III Z 1 KI-VO von biometrischen KI-Systemen, „soweit ihr Einsatz nach einschlägigem Unionsrecht oder nationalem Recht zugelassen ist“.

209 Vgl. die Beispiele in ErwGr 15 KI-VO sowie *Wendehorst* in *Martini/Wendehorst*, KI-VO, Art 3, Rz 238 (unter anderem „Unterschriftenidentifikation“). Zum Begriff der „speziellen technischen Verfahren“ siehe ebd., Rz 243 ff.

Was mit der „erstmaligen Identifizierung eines potenziellen Verdächtigen auf der Grundlage objektiver und nachprüfbarer Tatsachen, die in unmittelbarem Zusammenhang mit der Straftat stehen", gemeint ist, müsste wohl noch durch die Rechtsprechung konkretisiert werden. Beispielsweise könnte diese Formulierung bedeuten, dass die genutzte Referenzdatenbank ausschließlich aus solchen Daten besteht, die dem konkreten Ermittlungsakt entstammen, dass also nicht auf weitere Datenbestände zurückgegriffen wird. – Was ist jedoch die Rechtsfolge dieses Tatbestands? Einerseits könnte sie darin bestehen, dass in solchen Fällen das Erfordernis einer Genehmigung überhaupt wegfiele, andererseits könnte sie lediglich darin liegen, dass mit dem Antrag auf nachträgliche Genehmigung mehr als 48 Stunden gewartet werden darf. Im Interesse der Grundrechte, vor allem im eingriffsintensiven Bereich der Strafverfolgung, kann nur die letzte Variante die richtige sein, um der textuellen bzw. sprachlichen Unschärfe zu begegnen.[210]

Art 26 Abs 10 KI-VO bezieht sich, soweit bisher besprochen, zunächst nur auf die Suche nach – einer begangenen Straftat – **Verdächtigen oder Verurteilten**. Ermittlungen in Bezug auf andere Personen, beispielsweise Opfer, Zeugen und Vermisste, aber auch potenzielle „Gefährder",[211] sind davon jedoch nicht erfasst. Solche Ermittlungen mithilfe von Systemen zur nachträglichen biometrischen Identifizierung zu führen, ist auch nicht von den Verbotstatbeständen des Art 5 Abs 1 erfasst. Demnach handelt es sich dabei um „gewöhnliche" Hochrisiko-KI-Systeme im Bereich der Strafverfolgung, für deren Einsatz in einer konkreten Ermittlung die KI-VO kein eigenes Genehmigungsverfahren vorgibt. Im weiteren Verlauf des Normtextes wird es jedoch weitaus allgemeiner:

> „In keinem Fall darf ein solches Hochrisiko-KI-System zur nachträglichen biometrischen Fernidentifizierung zu Strafverfolgungszwecken in **nicht zielgerichteter Weise und ohne jeglichen Zusammenhang** mit einer Straftat, einem Strafverfahren, einer tatsächlichen und bestehenden oder tatsächlichen und vorhersehbaren Gefahr einer Straftat oder der Suche nach einer bestimmten vermissten Person verwendet werden. Es muss sichergestellt werden, dass die Strafverfolgungsbehörden keine ausschließlich auf der Grundlage der Ausgabe solcher Systeme zur nachträglichen biometrischen Fernidentifizierung beruhende **Entscheidung**, aus der sich eine **nachteilige Rechtsfolge für eine Person** ergibt, treffen."

210 So auch *I. Eisenberger* in *Martini/Wendehorst*, KI-VO, Art 26, Rz 55.

211 Der Ausdruck „einer Straftat verdächtigt" könnte natürlich auf den ersten Blick dahingehend verstanden werden, dass er auch Personen erfasse, die bloß der Begehung einer zukünftigen Straftat, also der Neigung zu einem „gefährlichen Angriff" verdächtigt werden. Dies liegt jedoch angesichts der sprachlichen Formulierung fern, außerdem weiß auch Art 5 Abs 1 lit h KI-VO über Echtzeit-Systeme sehr wohl zwischen faktisch begangenen Straftaten und erwarteten gefährlichen Angriffen zu unterscheiden.

Somit gilt für alle Systeme zur nachträglichen biometrischen FI ohne Unterschied ein „Verbot der anlasslosen Überwachung“.[212] Zudem enthält die Vorschrift – genauso wie diejenige über die biometrische Echtzeit-FI – das Recht, keiner nachteiligen Entscheidung unterworfen zu werden, die ausschließlich aus der Verwendung des KI-Systems resultiert. Auch hier ist die gründliche Einrichtung menschlicher Aufsicht und sorgfältige individuelle Aktenführung essenziell. Beispielsweise darf eine Festnahme nicht allein aufgrund einer biometrischen FI durch das KI-System erfolgen.[213] Im Sinne des ErwGr 95 dürfen die Systeme nur auf „verhältnismäßige, legitime und unbedingt erforderliche Weise eingesetzt werden und somit zielgerichtet sein, was die zu identifizierenden Personen, den Ort und den zeitlichen Anwendungsbereich betrifft, und auf einem geschlossenen Datensatz rechtmäßig erworbener Videoaufnahmen basieren“, falls überhaupt mit Videoaufnahmen gearbeitet wird. Sie dürfen auf keinen Fall zu „willkürlicher Überwachung“ führen.

Auch für die nachträgliche biometrische Fernidentifizierung gelten die besonderen Vorschriften über die Protokollierung nach Art 12 Abs 3 KI-VO.

Ein weiteres Erfordernis besteht darin, dass jede Verwendung des KI-Systems „in der einschlägigen **Polizeiakte dokumentiert**“ werden muss und dass diese Dokumentation der Marktüberwachungs- und der Datenschutzbehörde zur Verfügung zu stellen ist, wobei auch hier sensible operative Daten nicht davon betroffen sind. Ferner sind diesen Stellen Jahresberichte vorzulegen. „Die Berichte können eine Zusammenfassung sein, damit sie mehr als einen Einsatz abdecken.“ Eine Einbeziehung der Europäischen Kommission ist in Art 26 Abs 10 KI-VO nicht vorgesehen.

Es empfiehlt sich, sämtliche Informationen, die im Zuge des KI-Einsatzes zur biometrischen Fernidentifizierung anfallen, zunächst in einem eigenen „**Spurenakt**“ festzuhalten.[214] Aufgrund der allgemein bekannten Anfälligkeit solcher KI-Systeme für Fehlschlüsse, Verzerrungen usw. sollten solche Daten erst dann zum eigentlichen Ermittlungsakt genommen werden, wenn sie von menschlichen Organen hinreichend plausibel verifiziert werden konnten. Die Ermittlung kann auf diese Weise vor einer womöglich beträchtlichen Menge an „Ballast“ geschützt werden, weil sich das KI-generierte Material oft genug bei näherer Überprüfung oder auch schon auf den ersten Blick als hinfällig erweist. Nicht zuletzt liegt dieses Vorgehen im Interesse des Datenschutzes und unbeteiligter Dritter.[215]

212 *I. Eisenberger* in *Martini/Wendehorst*, KI-VO, Art 26, Rz 58 ff.
213 Vgl. ebd., Rz 60.
214 Zu diesem Begriff *Zerbes*, Spitzeln, Spähen, Spionieren, 152 ff, mit weiteren Nachweisen.
215 „Datenschutzrechtlich macht es einen entscheidenden Unterschied, ob die personenbezogenen Daten ein Durchlaufposten bleiben oder ob sie während des gesamten weiteren Strafverfahrens mitgezogen werden.“ (ebd., 153)

Es ist damit zu rechnen, dass im Bereich des sicherheitspolizeilichen **Erkennungsdienstes** nach §§ 67 ff SPG bereits mit Verfahren und Technologien gearbeitet wird,[216] welche für den Begriff des „Systems zur nachträglichen biometrischen Identifizierung" relevant sind. Da die Frage nach der Definition des KI-Begriffs noch nicht abschließend geklärt ist und somit dessen nähere Eingrenzung aussteht, sollte im Zweifel davon ausgegangen werden, dass sämtliche Technologien zur automatisierten biometrischen Fernidentifizierung KI-Systeme sind. Das bedeutet, dass solche erkennungsdienstlichen Anwendungen den Vorschriften der KI-VO unterfallen, mit den Polizeibehörden als Betreibern. Die polizeiliche Praxis muss daher gemäß Art 111 Abs 2 insofern bis zum 2. August 2030 den Vorgaben der Verordnung angeglichen werden. Für ein solches Vorgehen sollten jedenfalls nähere gesetzliche Vorschriften erlassen werden.[217] Dies gilt in besonders hohem Maß auch für den Fall, dass Personen auf **Videoaufnahmen nach § 54 Abs 5, 6, 7 und 8 SPG** mittels KI identifiziert werden sollen. Dort ist nach geltender Rechtslage überhaupt nicht von einer KI-Verwendung die Rede, daher kann sie – als ein gegenüber herkömmlichem Vorgehen intensivierter Grundrechtseingriff – derzeit nicht darauf gestützt werden.

4.4. Information betroffener Personen, Erläuterung der Entscheidungsfindung

Die KI-VO enthält in Art 26 Abs 11 und Art 86 zwei wichtige Vorschriften, welche den Betreiber dazu verpflichten, betroffene Personen zu einem gewissen Umfang über das KI-System und dessen Einsatz zu informieren. Gemäß Art 26 Abs 11 „informieren die Betreiber der in Anhang III aufgeführten Hochrisiko-KI-Systeme, die natürliche Personen betreffende Entscheidungen treffen oder bei solchen Entscheidungen Unterstützung leisten, die natürlichen Personen **darüber, dass sie der Verwendung des Hochrisiko-KI-Systems unterliegen**." Diese Information hat der Betreiber von sich aus, also ohne einen Antrag oder eine Nachfrage der betroffenen Person, zu erbringen. Sie „hat lediglich zu enthalten, dass ein konkretes System verwendet wurde"[218] oder verwendet wird. Art 86 Abs 1 geht jedoch deutlich weiter: Nach diesem Artikel haben betroffene Personen einer KI-unterstützten Entscheidung „das Recht, vom Betreiber eine klare und aussagekräftige Erläuterung zur **Rolle des KI-Systems im Entscheidungsprozess** und zu den **wichtigsten Elementen der getroffenen Entscheidung** zu erhalten." (*explainable AI*) ErwGr 171 setzt hinzu, die Erläuterung „sollte eine Grundlage bieten, auf der die betroffenen Personen ihre Rechte ausüben können."

216 Kritisch zu diesem Thema *Zerbes*, Spitzeln, Spähen, Spionieren, 285 ff.
217 § 75 SPG bietet bis heute keine Grundlage für einen Abgleich oder eine Analyse personenbezogener Daten mittels KI. Vgl. zur früheren Rechtslage *Hanel/Exenberger*, Automatisch ein Problem, juridikum 2021, 164 [169 f].
218 *I. Eisenberger* in *Martini/Wendehorst*, KI-VO, Art 26, Rz 76.

Die „Betroffenheit“ einer Person besteht nach dieser Vorschrift darin, dass die Entscheidung „rechtliche Auswirkungen“ auf sie hat „oder sie in ähnlicher Art erheblich auf eine Weise beeinträchtigt, die ihrer Ansicht nach ihre Gesundheit, ihre Sicherheit oder ihre Grundrechte beeinträchtigt“. „Eine Beeinträchtigung der Grundrechte der Person könnte beispielsweise durch Anknüpfung an diskriminierende Unterscheidungskriterien und Gruppenzugehörigkeiten erfolgen.“[219] Die Verpflichtung des Betreibers zur Erläuterung der Entscheidungsfindung ist jedoch an einen Antrag der betroffenen Person bzw. an eine „aktive Aufforderung“ gebunden.[220] Als „Entscheidung“ im Sinne von Art 86 KI-VO gilt nicht nur „die letztliche 'Haupt-Entscheidung'“, sondern auch „vorgelagerte bzw. untergeordnete 'Vor-Entscheidungen'“ sind darunter zu verstehen.[221] Die Einschätzung eines Risikos,[222] welches mutmaßlich von einer Gruppierung auf einem Internetforum ausgeht, kann eine solche Vorentscheidung darstellen; ferner die Identifikation einer Person in einem Video mittels KI-gestützter Bild- oder Spracherkennung; gleichermaßen die Annahme, Person A stehe in regelmäßigem Kontakt zu Person B, gegebenenfalls zu bestimmten Zeiten, an bestimmten Orten, über bestimmte Medien bzw. Kommunikationskanäle, zur Besprechung bestimmter Themen (*social network analysis*); oder auch die Bewertung einer Information als relevant für die Frage, welche Stimmung in einem gegebenen Kontext vorherrscht (*sentiment analysis*).

Die Erläuterung, welche sich grob gesprochen auf die „Rolle des KI-Systems im Entscheidungsprozess“ und auf die „wichtigsten Elemente der getroffenen Entscheidung“ bezieht, sollte die einbezogenen **Faktoren**, deren **Gewichtung** sowie die durch das System bzw. daran anknüpfend durch einen Menschen gezogenen **Schlüsse** umfassen. Dabei dürfte den automatisch erzeugten Protokollen gemäß Art 12 KI-VO praktische Bedeutung zukommen. Schließlich wird auch zumindest in Grundzügen über die involvierte **Logik** der Entscheidungsfindung aufzuklären sein. Dazu muss es nicht immer nötig sein, den Algorithmus selbst offenzulegen,[223] insbesondere wenn er für die betroffene Person ohnehin völlig unverständlich ist[224] und die Offenlegung etwa den staatlichen Schutz von Rechtsgütern beeinträchtigen könnte. Vielmehr kommt es darauf an, die Entscheidung für die betroffene Person **nachvollziehbar** zu machen.[225] Die DSGVO verlangt, dass in Fällen der „automati-

219 *Hartmann* in *Martini/Wendehorst*, KI-VO, Art 86, Rz 13.
220 Ebd., Rz 14.
221 Ebd., Rz 15.
222 Zu „Score“- bzw. Wahrscheinlichkeitswerten über das zukünftige Verhalten oder Möglichkeiten von Personen aktuell EuGH 07.12.2023, C-634/21 (*SCHUFA Holding [Scoring]*); VwGH 21.12.2023, Ra 2021/04/0010.
223 Vgl. zu dieser Thematik *Dürager*, Künstliche Intelligenz – eine besondere Art des Profiling nach der DSGVO, JBD 2019, 375 [390 f], mit Bezug auf geistiges Eigentum und Geschäftsgeheimnisse.
„Die betroffene Person ist ungeachtet dessen jedoch so zu informieren, dass sie die an der Bewertung beteiligten Parameter versteht und am Ende feststellen kann, welche Aspekte ihrer Person oder ihres Verhaltens bei der Erstellung des Profils oder der Berechnung eines Score-Wertes berücksichtigt wurden.“ (ebd., 391)
224 Vgl. *Illibauer* in *Knyrim,* DatKomm Art 13 DSGVO (Stand 1.7.2024, rdb.at), Rz 55, mwN; *Zavadil*, Der besondere Auskunftsanspruch über die involvierte Logik einer Datenverarbeitung, Dako 2020, 55 [56].
225 Vgl. *Art-29-Datenschutzgruppe*, Leitlinien zu automatisierten Entscheidungen im Einzelfall einschließlich Profiling

sierten Entscheidungsfindung einschließlich Profiling“ im Sinne des Art 22 aussagekräftige Information über deren Bestehen, „über die involvierte Logik sowie die Tragweite und die angestrebten Auswirkungen einer derartigen Verarbeitung für die betroffene Person“ erteilt wird.[226] Speziell zu Art 86 KI-VO wird ausgeführt, dass „zunächst das Zustandekommen und sodann die wertende Bedeutung der Daten bzw. Ausgaben offengelegt werden“ und dass zudem darauf hinzuweisen sein wird, „welcher Anteil der Entscheidung durch das System determiniert wurde und welcher Spielraum für Überprüfung und Korrektur dem Betreiber abstrakt verbleibt und ob und wie dieser im Einzelfall genutzt wurde.“[227]

Nun beinhaltet sowohl Art 86 Abs 2 als auch Art 26 Abs 11 KI-VO eine Klausel, welche eine Einschränkung der Information aus Gründen des Datenschutzrechts zulässt. Während Art 86 Abs 2 pauschal von „Ausnahmen von oder Beschränkungen der Pflicht […] aus dem Unionsrecht oder dem nationalen Recht im Einklang mit dem Unionsrecht“ spricht,[228] enthält Art 26 Abs 11 für den Bereich **Strafverfolgung** einen Verweis auf Art 13 DSRL-PJ,[229] dessen Abs 3 (gesetzliche) Einschränkungen der Unterrichtung betroffener Personen über Datenverarbeitungen erlaubt. Demnach darf das nationale Recht vorsehen, dass „die Unterrichtung der betroffenen Person […] soweit und so lange aufgeschoben, eingeschränkt oder unterlassen werden kann, wie diese Maßnahme in einer demokratischen Gesellschaft erforderlich und verhältnismäßig ist und sofern den Grundrechten und den berechtigten Interessen der betroffenen natürlichen Person Rechnung getragen wird:

> a) zur Gewährleistung, dass behördliche oder gerichtliche Untersuchungen, Ermittlungen oder Verfahren nicht behindert werden,
> b) zur Gewährleistung, dass die Verhütung, Aufdeckung, Ermittlung oder Verfolgung von Straftaten oder die Strafvollstreckung nicht beeinträchtigt werden,
> c) zum Schutz der öffentlichen Sicherheit,
> d) zum Schutz der nationalen Sicherheit,
> e) zum Schutz der Rechte und Freiheiten anderer.“

für die Zwecke der Verordnung 2016/679 (WP251rev.01, Fassung 06.02.2018), 28.

226 Art 13 Abs 2 lit f, Art 14 Abs 2 lit g, Art 15 Abs 1 lit h DSGVO.

227 *Hartmann* in *Martini/Wendehorst*, KI-VO, Art 86, Rz 15. Wie besprochen gibt freilich die KI-VO an verschiedenen Stellen vor, dass der genannte Spielraum für menschliche Organe im Bereich Strafverfolgung durch die Verwendung des KI-Systems letztlich überhaupt nicht beschränkt werden darf; so ist es auch verfassungsrechtlich betrachtet notwendig, dass die Verantwortung für Ermittlungen ausschließlich bei den dazu bestellten – menschlichen – Organen liegt.

228 Solche Einschränkungen des Rechts auf Erläuterung müssen freilich immer **verhältnismäßig** sein, weil sie – zumindest mittelbar – einen Eingriff in Grundrechte, insbesondere auf ein faires Verfahren gemäß Art 6 EMRK, darstellen.

229 Zur Umsetzung in Österreich siehe § 43 Abs 4 und § 44 Abs 2 DSG. Vgl. § 12 Abs 4 SNG.

Wird jedoch nach der StPO wegen einer strafbaren Handlung ermittelt oder liegt im Sinne des Art 6 EMRK ein sonstiges Verfahren über eine „strafrechtliche Anklage“ oder auch über „zivilrechtliche Ansprüche und Verpflichtungen“ vor – wozu auch Streitigkeiten öffentlicher Bediensteter zählen[230] –, dann sind die Grundsätze eines fairen Verfahrens nach diesem Grundrechtsartikel einzuhalten. Zentral dafür sind die Grundsätze des kontradiktorischen Verfahrens und der Waffengleichheit, welche das Recht auf Akteneinsicht umfasst.[231] Jedoch ist es nach der Rechtsprechung des EGMR möglich, „dass übergeordneten nationalen Interessen Vorrang eingeräumt wird, wenn einer Partei ein vollständig kontradiktorisches Verfahren verweigert wird. Den Vertragsstaaten kommt in diesem Bereich ein gewisser Ermessensspielraum zu.“[232]

> „Der [Gerichtshof] wiederholt zudem, dass der **Anspruch auf Offenlegung von relevanten Beweisen** ebenfalls kein absoluter ist. In Strafverfahren hat er festgestellt, dass konkurrierende Interessen vorliegen können, die gegen die Rechte der Verfahrenspartei abgewogen werden müssen, wie die **nationale Sicherheit** oder die Notwendigkeit, **Zeugen** zu schützen, denen Repressalien drohen, oder **polizeiliche Verbrechensermittlungsmethoden geheim zu halten**. Jedoch sind unter Art. 6 Abs. 1 EMRK nur Maßnahmen zur Einschränkung der Rechte einer Verfahrenspartei erlaubt, die nicht das **Wesen dieser Rechte** beeinträchtigen. Damit dies der Fall ist, müssen Schwierigkeiten, die einer bf. Partei durch eine Beschränkung ihrer Rechte verursacht werden, ausreichend durch das von den Gerichten verfolgte Verfahren **kompensiert** werden.“[233]

Werden also bestimmte Dokumente oder Information wie beispielsweise der Algorithmus oder andere Komponenten des an einer Entscheidung beteiligten KI-Systems von der Akteneinsicht ausgenommen, muss diese Beeinträchtigung der Verfahrensrechte der betroffenen Personen durch andere Schutzvorkehrungen ausgeglichen werden. Das Verfahren ist somit als Ganzes danach zu beurteilen, ob es den Parteirechten gemäß Art 6 EMRK genügt.[234] Wirkt sich eine Entscheidung eines KI-Systems maßgeblich auf das Strafverfahren oder gar auf das gerichtliche Urteil aus, stellt es freilich einen besonders gravierenden Eingriff in das Recht auf ein faires Verfahren – darunter das **Recht auf Begründung**[235] – dar, wenn die Funktionsweise des KI-Systems nicht genau genug erläutert wird oder die einbezogenen Daten nicht angeführt werden. Ist dennoch die Geheimhaltung der In-

230 Dazu *Muzak,* B-VG⁶ Art 6 MRK (Stand 1.10.2020, rdb.at), Rz 3, 6, mwN; EGMR 19.9.2017, Bsw 35.289/11 (Große Kammer), *Regner gegen Tschechische Republik*, NLMR 2017, 425.
231 Vgl. *Muzak,* B-VG⁶ Art 6 MRK (Stand 1.10.2020, rdb.at), Rz 40.
232 EGMR, *Regner*, Absatz 147. In der Entscheidung geht es um **geheimdienstliche** Informationen.
233 Ebd., Absatz 148.
234 Vgl. ebd., Absatz 149 ff.
235 Vgl. speziell zu Art 6 EMRK *Kischel*, Die Begründung (2003), 139 ff.

formation gegenüber der betroffenen Person aus den vom EGMR zugelassenen Gründen unerlässlich, muss das entscheidende Gericht selbst auf jeden Fall in vollem Umfang Zugang zu den relevanten Informationen haben und diese unabhängig überprüfen.[236] Im Zweifel sollte jedoch die betroffene Person, „wenn auch nur summarisch",[237] stets über alle für das Verfahren wesentlichen Punkte aufgeklärt werden.

In diesem Zusammenhang ist ferner von Bedeutung, dass Art 11 DSRL-PJ eine „automatisierte Entscheidungsfindung im Einzelfall" einschließlich Profiling nur dann erlaubt, wenn das Recht des Mitgliedstaats „geeignete Garantien für die Rechte und Freiheiten der betroffenen Person bietet". Diese können unter anderem darin bestehen, dass die Person über die Logik, die Faktoren und deren Gewichtung informiert wird oder dass sie die Entscheidung anfechten kann. „Die Auswirkungen des Einsatzes von KI-Instrumenten auf die **Verteidigungsrechte** von Verdächtigen sollten nicht außer Acht gelassen werden, insbesondere nicht die Schwierigkeit, aussagekräftige Informationen über die Funktionsweise solcher Systeme zu erhalten, und die daraus resultierende Schwierigkeit einer **gerichtlichen Anfechtung** ihrer Ergebnisse, insbesondere durch natürliche Personen, gegen die ermittelt wird."[238] Letztlich darf im Bereich Strafverfolgung eine Entscheidung eines KI-Systems einer Entscheidung durch ein menschliches Organ ohnehin nicht vorgreifen, sodass dieses sein Handeln selbst verantworten und mit dergleichen Ausgaben immer kritisch umgehen muss.

Art 86 Abs 3 KI-VO dürfte im vorliegenden Themenbereich irrelevant sein, weil das Unionsrecht – anders als bei Art 22 DSGVO für Verarbeitungen außerhalb von Polizei und Justiz[239] – keine weitere Rechtsgrundlage für die Erläuterung der Entscheidungsfindung im Einzelfall für den Bereich Strafverfolgung kennt.

Nebenbei ist noch zu erwähnen, dass Betreiber eines KI-Systems zur Emotionserkennung oder zur biometrischen Kategorisierung nach Art 50 Abs 3 KI-VO die betroffene Person grundsätzlich über den Betrieb des Systems informieren sollen; dies gilt jedoch nicht im Bereich Strafverfolgung, „sofern geeignete Schutzvorkehrungen für die **Rechte und Freiheiten Dritter** bestehen."

236 Vgl. EGMR, *Regner*, Absatz 152 ff; EuGH 13. 9. 2018, C-358/16, (*UBS Europe ua*); EuGH 14. 2. 2008, C-450/06 (*Varec*); *Kröll* in *Holoubek/Lienbacher,* GRC-Kommentar² Art 47 (Stand 1.4.2019, rdb.at), Rz 85.
237 EGMR, *Regner*, Absatz 160; vgl. ebd., Absatz 153.
238 ErwGr 59 KI-VO.
239 Vgl. *Hartmann* in *Martini/Wendehorst*, KI-VO, Art 86, Rz 17 ff.

Exkurs: KI-Einsatz und Werturteile

Eine begriffliche Unterscheidung, welche für den Einsatz von KI-Systemen in Ermittlungen des Bereichs Strafverfolgung von Bedeutung sein wird, ist die zwischen **Tatsachen- und Werturteilen bzw. „persönlichen Einschätzungen“**. Dafür bestimmt bereits § 37 Abs 5 DSG in Umsetzung des Art 7 Abs 1 DSRL-PJ: „Soweit möglich ist zwischen faktenbasierten und auf persönlichen Einschätzungen beruhenden personenbezogenen Daten zu unterscheiden. Auf persönlichen Einschätzungen beruhende personenbezogene Daten sind entsprechend zu **kennzeichnen** und können mit einer **Begründung** versehen werden, welche die Nachvollziehbarkeit der Einschätzung ermöglicht.“ Dies muss umso mehr gelten, wenn betreffende Daten von einem KI-System weiter verarbeitet werden, wenn daraus insbesondere in automatisierter Weise **Schlüsse** gezogen werden. Im Rahmen der konkreten **Aktenführung** sollten daher als KI-Ausgaben generierte Informationen, die in irgendeiner Art auf Werturteilen oder persönlichen Einschätzungen beruhen, ebenfalls als solche ausgewiesen werden. Faktenbasierte Urteile sind beispielsweise: „F hat grüne Augen“, „D postet jeden Tag auf dem Internetforum x“, „P hat letztes Jahr mehrere Reisen gemeinsam mit bestimmten Personen unternommen“; dagegen sind persönliche Einschätzungen oder Werturteile: „J verhält sich gleichgültig gegenüber der Gesellschaft“, „T hat womöglich einen extremistischen Hintergrund“, „K ist einer Straftat verdächtig“, „G könnte tendenziell eine Gefahr für die Sicherheit darstellen“. Ist also eine Information der letzteren Kategorie – etwa als Eingabedatum – in den Prämissen eines vom KI-System gezogenen Schlusses enthalten, so hat auch dessen Konklusion als eine persönliche Einschätzung oder als Werturteil zu gelten, ist damit solcherart zu kennzeichnen.

Exkurs: Datenverordnung

Art 14 ff der DatenVO[240] („Data Act“) regeln die „Bereitstellung von Daten für öffentliche Stellen, die Kommission, die Europäische Zentralbank und Einrichtungen der Union wegen **außergewöhnlicher Notwendigkeit**“. Nach Art 15 sind „Dateninhaber“ dazu verpflichtet, den genannten Stellen Daten bereitzustellen, sofern letztere „den Nachweis dafür“ erbringen, „dass im Hinblick auf die Erfüllung ihrer rechtlichen Aufgaben im öffentlichen Interesse die außergewöhnliche Notwendigkeit der Nutzung bestimmter Daten“ besteht. In Art 2 Z 13 wird der „**Dateninhaber**“ definiert als „eine natürliche oder juristische Person, die nach dieser Verordnung, nach geltendem Unionsrecht oder

240 Verordnung (EU) 2023/2854 des Europäischen Parlaments und des Rates vom 13. Dezember 2023 über harmonisierte Vorschriften für einen fairen Datenzugang und eine faire Datennutzung sowie zur Änderung der Verordnung (EU) 2017/2394 und der Richtlinie (EU) 2020/1828 (Datenverordnung) (ABl. L, 2023/2854, 22.12.2023, ELI: http://data.europa.eu/eli/reg/2023/2854/oj).

nach nationalen Rechtsvorschriften zur Umsetzung des Unionsrechts berechtigt oder verpflichtet ist, Daten – soweit vertraglich vereinbart, auch Produktdaten oder verbundene Dienstdaten – zu nutzen und bereitzustellen, die sie während der Erbringung eines verbundenen Dienstes abgerufen oder generiert hat".[241] Dabei sind Daten nach Z 1 zu verstehen als „jede digitale Darstellung von Handlungen, Tatsachen oder Informationen sowie jede Zusammenstellung solcher Handlungen, Tatsachen oder Informationen auch in Form von Ton-, Bild- oder audiovisuellem Material". Der Begriff des „**verbundenen Dienstes**" bezeichnet laut Z 6 einen „digitalen Dienst, bei dem es sich nicht um einen elektronischen Kommunikationsdienst handelt, – einschließlich Software –, der zum Zeitpunkt des Kaufs, der Miete oder des Leasings so mit dem Produkt verbunden ist, dass das vernetzte Produkt ohne ihn eine oder mehrere seiner Funktionen nicht ausführen könnte oder der anschließend vom Hersteller oder einem Dritten mit dem Produkt verbunden wird, um die Funktionen des vernetzten Produkts zu ergänzen, zu aktualisieren oder anzupassen". Gemäß Z 5 ist das „**vernetzte Produkt**" ein „Gegenstand, der Daten über seine Nutzung oder Umgebung erlangt, generiert oder erhebt und der Produktdaten über einen elektronischen Kommunikationsdienst, eine physische Verbindung oder einen geräteinternen Zugang übermitteln kann und dessen Hauptfunktion nicht die Speicherung, Verarbeitung oder Übertragung von Daten im Namen einer anderen Partei – außer dem Nutzer – ist". Es geht somit um den Themenkreis, der als *Internet of Things (IoT)* bezeichnet wird.

Art 15 DatenVO bestimmte, dass eine „außergewöhnliche Notwendigkeit der Nutzung bestimmter Daten […] zeitlich befristet und im Umfang begrenzt" ist und nur dann als gegeben gilt, wenn entweder a) „die verlangten Daten zur Bewältigung eines **öffentlichen Notstands** erforderlich sind und die öffentliche Stelle […] diese Daten unter gleichwertigen Bedingungen auf andere Weise nicht rechtzeitig und wirksam beschaffen kann", oder b) es sich nur um nicht-personenbezogene Daten handelt, welche zur Erfüllung einer rechtlich vorgesehenen Aufgabe öffentlichen Interesses – beispielsweise amtliche Statistik – notwendig sind. Für b) werden in diesem Artikel weitere Bedingungen und Einschränkungen festgelegt. Der „öffentliche Notstand" wird in Art 2 Z 29 definiert als „eine zeitlich begrenzte Ausnahmesituation – wie etwa Notfälle im Bereich der öffentlichen Gesundheit, Notfälle infolge von Naturkatastrophen sowie von Menschen verursachte Katastrophen größeren Ausmaßes, einschließlich schwerer Cybersicherheitsvorfälle –, die sich negativ auf die Bevölkerung der Union oder eines Mitgliedstaats bzw. eines Teils davon auswirkt, das Risiko schwerwiegender und dauerhafter Folgen für die Lebensbedingungen, die wirtschaftliche Stabilität oder die finanzielle Stabilität oder die Gefahr einer erheblichen und unmittelbaren Beeinträchtigung wirt-

241 Für die Definitionen von „Produktdaten" und „verbundene Dienstdaten" siehe Art 2 Z 15 und 16 DatenVO.

schaftlicher Vermögenswerte in der Union oder in dem betroffenen Mitgliedstaat birgt und die nach den einschlägigen Verfahren des Unionsrechts oder des nationalen Rechts festgestellt und amtlich ausgerufen wurde“. Der Verordnungsgeber hat somit bei der Bestimmung dieses Begriffs jedenfalls keinen besonderen Fokus auf die Aufgaben der Sicherheitspolizei und der Strafverfolgung gelegt.

Ganz unabhängig davon nimmt schließlich Art 16 Abs 2 DatenVO Behörden, „die Tätigkeiten zur Verhütung, Ermittlung, Aufdeckung oder Verfolgung von Straftaten oder Ordnungswidrigkeiten oder der Strafvollstreckung durchführen“, überhaupt vom Anwendungsbereich sämtlicher Vorschriften des Kapitels V (Art 14 ff) über die besprochene Bereitstellung von Daten an Behörden durch Dateninhaber aus. „Dieses Kapitel berührt nicht das anwendbare Unionsrecht und das anwendbare nationale Recht über die Verhütung, Ermittlung, Aufdeckung oder Verfolgung von Straftaten oder Ordnungswidrigkeiten oder über die Vollstreckung von Strafen oder verwaltungsrechtlichen Sanktionen“. Das bedeutet, dass ein Verlangen zur Herausgabe von Daten derzeit weder im Rahmen konkreter Ermittlungen – wie in § 138 StPO, § 53 Abs 3c SPG oder § 11 Abs 2 SNG – noch für Zwecke der Erstellung amtlicher Statistik – damit verwandt die anonymisierte Trendforschung – auf die DatenVO gestützt werden kann.

Literatur

Art-29-Datenschutzgruppe, Leitlinien zu automatisierten Entscheidungen im Einzelfall einschließlich Profiling für die Zwecke der Verordnung 2016/679 (WP251rev.01, Fassung 06.02.2018)

Ashley, Artificial Intelligence and Legal Analytics. New Tools for Law Practice in the Digital Age (2017)

[AutorIn] in *Bertel/Venier* (Hrsg), StPO: Kommentar² (2022)

[AutorIn] in *Birklbauer/Haumer/Nimmervoll/Wess* (Hrsg), StPO – Linzer Kommentar zur Strafprozessordnung (2020)

Bußjäger, "Weisungsfreie Verwaltung und Beleihung", ÖZW 2022, 190

Dürager, Künstliche Intelligenz – eine besondere Art des Profiling nach der DSGVO, JBD 2019, 375

Figl/Müller, Verfassungsschutz neu: Ist der Nachrichtendienst (weiterhin) Kriminalpolizei?, JSt 2022, 135

[AutorIn] in *Fuchs/Ratz*, WK StPO (rdb.at)

Gölly/Kaiser, Strafrecht grenzenlos? Von Grenzen und ihren Überschreitungen, JSt 2024, 298

Hanel/Exenberger, Automatisch ein Problem, juridikum 2021, 164

Hart, The Concept of Law (1997², 1961)

Heißl, Darf die Polizei googeln? Der Gebrauch des World Wide Web in sicherheitspolizeilichen Ermittlungen, ZfV 2016, 158

--- Überwachungen und Ermittlungen im Internet (2017)

[AutorIn] in *Heißl/Figl* (Hrsg), SNG² (2023)

Helminger, Datenschutzrechtliche Herausforderungen bei der Verwendung von Trainingsdaten, EALR 2022, 46

Holoubek, Grundrechtsschutz vor neuen Herausforderungen, 21. ÖJT Band I/1 (2022)

[AutorIn] in *Holoubek/Lienbacher,* GRC-Kommentar² (rdb.at)

Hönigswald, „Begriff“ und „Interesse“. Eine methodologische Skizze. Archiv für civilistische Praxis (AcP) 1934, 1

[AutorIn] in *Jaeger/Stöger* (Hrsg) (rdb.at)

Kastelitz/Tschohl/Hötzendorfer, (Datenschutz-)Rechtliche Aspekte der polizeilichen Verarbeitung von Videomassendaten, JBD 2019, 327

Kelsen, Reine Rechtslehre (2017, 1960²)

Kieseberg/Kaltenbrunner/Temper/Tjoa, Beschaffung sicherer AI-Systeme – Herausforderungen & Praktischer Leitfaden (http://ffhoarep.fh-ooe.at/bitstream/123456789/1726/1/FFH2023_244.pdf, Abruf 17.11.2024)

Kischel, Die Begründung. Zur Erläuterung staatlicher Entscheidungen gegenüber dem Bürger (2003)

Klamert, EU-Recht (2015)

Klaushofer, Strukturfragen der Rechtsschutzbeauftragten (2012)

[AutorIn] in *Knyrim*, DatKomm (rdb.at)

Lachmayer, Rechtsstaatliche Grenzen polizeilicher Überwachungsbefugnisse, JBÖR 2020, 105

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

[AutorIn] in *Martini/Wendehorst*, KI-VO. Verordnung über künstliche Intelligenz. Kommentar (2024)

Mayrhofer, Rechtliche Herausforderungen und Grenzen des Einsatzes von Assistenzsystemen in der Gerichtsbarkeit, RZ 2024, 27

Mayrhofer/Parycek, Digitalisierung des Rechts – Herausforderungen und Voraussetzungen, 21. ÖJT Band IV/1 (2022)

Muzak, B-VG⁶ (rdb.at)

Oskierski, Schadenersatz im Europäischen Recht. Eine vergleichende Untersuchung des Acquis Communitaire und der EMRK (2010)

Palmer, European contractual regimes: The contemporary approaches. In: *Palmer* (ed.), The Recovery of Non-Pecuniary Loss in European Contract Law (2015), 95

Pfister, Does the Criticality Pyramid of the Proposed European AI Act Have Blind Spots? A Comparative Analysis of EU and US Law (https://law.stanford.edu/publications/no-117-does-the-criticality-pyramid-of-the-proposed-european-ai-act-have-blind-spots-a-comparative-analysis-of-eu-and-us-law/, Abruf 05.07.2024)

Pollirer, Checkliste Videoüberwachung nach der EDPB-Leitlinie 3/2019, Dako 2021, 12

--- Checkliste AI Act, Dako 2024, 62

Rachinger/Voithofer/Kettemann, Jenseits der Daten-Monokulturen: Zur Notwendigkeit diversitäts-sensibler Trainingsdaten, jusIT 2024, 43

Raz, The Authority of Law. Essays on Law and Morality (1979)

Reindl-Krauskopf, Grundrechtliche und strafprozessrechtliche Grenzen: Welche rote Linien zieht das Recht?, AnwBl 2024, 35

Rothmund-Burgwall, Einsatz von KI als neuartige Datenverarbeitung?, Dako 2024, 55

Salimi, Gefährliche Gruppierungen als Sicherheitsbedrohung. Grenzen und Alternativen im Kampf gegen Terror und organisierte Kriminalität durch Strukturermittlungen und Strafe (2022)

Schemmel, Grundrechte- und Datenschutz-Folgenabschätzung – zwei Seiten einer Medaille des Daten-Risikomanagements, CB 2024, 321

[AutorIn] in *Schmölzer/Mühlbacher* (Hrsg), StPO Kommentar, Band 1 Ermittlungsverfahren² (2021)

Schürmann, KI und Recht: Die Datenschutz-Folgenabschätzung und die Entwicklung von KI-Systemen, DSB 2021, 320

Schwaiger, Biometrische Gesichtserkennung, JBD 2016, 193

Schweighofer, Rechtsinformatik und Wissensrepräsentation. Automatische Textanalyse im Völkerrecht und Europarecht (1999)

Thiele/Wagner, Praxiskommentar zum Datenschutzgesetz (DSG)² (rdb.at)

Unabhängige Hochrangige Expertengruppe für künstliche Intelligenz, Ethik-Leitlinien für eine vertrauenswürdige KI (2019, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, Abruf 02.07.2024)

--- Eine Definition der KI: Wichtigste Fähigkeiten und Wissensgebiete (2019, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, Abruf 02.07.2024)

Wendehorst/Nessler/Aufreiter/Aichinger, Der Begriff des „KI-Systems" unter der neuen KI-VO, MMR 2024, 605

Wurmnest, Grundzüge eines europäischen Haftungsrechts (2003)

Zanol, Über den Schutzbereich öffentlich verfügbarer Daten im Anwendungsbereich der europäischen Datenschutz-Grundverordnung (DSGVO) (Dissertation, Uni Wien, unveröffentlicht)

Zavadil, Der besondere Auskunftsanspruch über die involvierte Logik einer Datenverarbeitung, Dako 2020, 55

Zerbes, Spitzeln, Spähen, Spionieren. Sprengung strafprozessualer Grenzen durch geheime Zugriffe auf Kommunikation (2011)