datenrecht.ch

AI Act: FAQ zur KI-Verordnung

Wichtige Fragen und vorläufige Antworten

Dr. David Vasella, Rechtsanwalt, CIPP/E, CIPM, FIP, Walder Wyss AG, Zürich

Version 1.0, 22. September 2024

Der Autor dankt für wertvolle Hinweise Amina Chammah, Lena Götzinger, Hannes Meyle und Kento Reutimann (alle Walder Wyss), und für fruchtbare Diskussionen David Rosenthal (Vischer).

Für Hinweise auf Fehler an david.vasella@walderwyss.com oder hello@datenrecht.ch sind wir dankbar.

Grundlagen

Grundfragen

Rollen

Verbotene und hochriskante Anwendungen

Kernpflichten bei HRAIS

GPAI

AIS im Betrieb

Sonderfragen

Sanktionen und Governance

Ergänzungsfragen

Grundlagen

1 Welche Begriffe werden in diesen FAQ verwendet?

Diese FAQ verwenden – neben den gesetzlich definierten Begriffen (→ 8) – die folgenden Abkürzungen:

AI	Künstliche Intelligenz
AIA	AI Act (KI-Verordnung). Verweisungen auf Artikel ohne andere Angaben beziehen sich jeweils auf den AIA
AIS	AI-System (KI-System)
FOSS	Free and Open-Source Software (freie und quelloffene Lizenz)
GPAI	General-Purpose AI (KI mit allgemeinem Verwendungszweck)
GPAIM	General-Purpose AI Model (KI-Modell mit allgemeinem Verwendungszweck)
GPAIS	General-Purpose AI System (KI-System mit allgemeinem Verwendungszweck)
HRAIS	High-Risk AIS (KI-System mit hohen Risiken)
QMS	Qualitätsmanagement-System
RMS	Risikomanagement-System

2 Was ist der AIA?

Die "Verordnung (EU) 2024/1689 vom 13. Juni 2024 zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz und zur Änderung [...]", die Verordnung über künstliche Intelligenz, KI-Verordnung, AI Act oder AIA) ist der umfassende regulatorische Rahmen, mit dem die EU (oder der EWR, der AIA ist von EWR-Relevanz) den Einsatz von KI-Systemen (AI-Systeme, AIS) regelt.

Eine englischsprachige Online-Fassung des AIA mit einer nicht verbindlichen Zuordnung der Erwägungsgründe findet sich bei datenrecht, ebenso wie eine PDF-Version.

Es ist eine Verordnung, also wie die DSGVO direkt anwendbar. Die zuständigen Behörden werden allerdings einige Punkte konkretisieren und ändern können (→ 51).

In der Sache definiert der AIA zuerst seinen sachlich und räumlich-persönlichen Anwendungsbereich und legt Regeln für die Entwicklung und den Einsatz von AIA fest, vor allem für "Hochrisiko-KI-Systeme" ("High-Risk AI Systems", "**HRAIS**") und für AIS mit allgemeiner Zwecksetzung (also use case-agnostische, breit einsetzbare AIS – sog. "General-Purpose AI", "**GPAI**" → 39). Bestimmte besonders unerwünschte Praktiken (Use Cases) werden zudem verboten (→ 27).

3 Wo finde ich weitere Informationen zum AI Act?

Die wissenschaftliche Aufarbeitung des AI Act ist im Gange, steht aber noch am Anfang. Aus der allgemeinen schweizerischen Literatur sei (unvollständig) auf folgende Beiträge verwiesen:

- **Rosenthal**, Der EU AI Act – Verordnung über künstliche Intelligenz, Jusletter vom 5. August 2024 (https://dtn.re/tLrdFm)
- **Arioli**, Risikomanagement nach der EU-Verordnung über Künstliche Intelligenz, Jusletter IT vom 4. Juli 2024 (https://dtn.re/7iE4zb)
- **Houdrouge/Kruglak**, Are Swiss data protection rules ready for AI?, Jusletter vom 27. November 2023 (https://dtn.re/KvghSt)
- **Müller**, Der Artificial Intelligence Act der EU: Ein risikobasierter Ansatz zur Regulierung von Künstlicher Intelligenz, EuZ 1/2022 (https://dtn.re/PafzEb)

Spezialliteratur findet sich insbesondere zu **urheberrechtlichen** Fragen im Zusammenhang mit generativer AI (bspw. Thouvenin/Picht, AI & IP: Empfehlungen für Rechtsetzung, Rechtsanwendung und Forschung zu den Herausforderungen an den Schnittstellen von [AI und IP], sic! 2023, 507 ff.), zu **Haftungsfragen** (bspw. Quadroni, Künstliche Intelligenz – praktische Haftungsfragen, HAVE 2021, 345 ff.), zu **arbeitsrechtlichen** Themen (bspw. Wildhaber, Künstliche Intelligenz und Mitwirkung am Arbeitsplatz, ARV 2024 1 ff.).

Weitere Informationen finden sich laufend auf www.datenrecht.ch und auf dem Blog von Vischer (https://dtn.re/BAG7II).

Aus der ausländischen juristischen Literatur sei insbesondere auf folgende Werke verwiesen:

- **Voigt/Hullen**, Handbuch KI-Verordnung FAQ zum EU AI Act, 2024 (Kindle E-Book: https://dtn.re/bIwQg3)
- **Wendt/Wendt**, Das neue Recht der Künstlichen Intelligenz, 2024 (Kindle E-Book: https://dtn.re/kFmWjk)

Aus der nicht oder nicht vorwiegend juristischen Literatur sei verwiesen auf:

- **Gasser/Mayer-Schönberger**, Guardrails: Guiding Human Decisions in the Age of AI, 2024, eine Diskussion von Rahmenbedingungen (Gesetze, Normen und Technologien) für Entscheidungen, der Herausforderungen digitaler Entscheidungen und möglicher Prinzipien für Leitplanken (Kindle E-Book: https://dtn.re/nYx3pm)
- **Strümke**, Künstliche Intelligenz (Kindle E-Book: https://dtn.re/eOI7vU); eine recht umfassende und lesbare Einführung zur Geschichte des Gebiets, technischen Fragen, Risiken und Schwachpunkten und Spekulationen zur weiteren Entwicklung.

4 Wie verlief die Verhandlung des AI Act?

Die Europäische Kommission hatte ihren Vorschlag am **21. April 2021** vorgelegt (Vorschlag der Europäischen Kommission vom 21. April 2021, https://dtn.re/vcdOAU). Der Rat der EU legte seine Position am **6. Dezember 2022** fest (Position des Rates der EU vom 6. Dezember 2022, https://dtn.re/Z9KtSp). Diskutiert wurden damals u.a. Ausnahmen für KMU, die Durchsetzung und Sanktionen und erneut der Einbezug von GPAI. Am **14. Juni 2023** hatte das Europäische Parlament seine Position zum AIA festgelegt (Position des Europäischen Parlaments vom 14. Juni 2023, https://dtn.re/JSQJtF), wobei vor allem strengere Vorschriften zur Transparenz und Nachvollziehbarkeit ein Anliegen war. Bereits intensiv diskutiert war damals die Regulierung von AI-Modellen, die sich für einen breiten Einsatz eignen ("General-Purpose AI Model", GPAIM; damals öfter als "Foundation Model" bezeichnet).

In den anschliessenden **Trilogverhandlungen**, dem informellen Verhandlungsverfahren, bei dem Vertreter des Parlaments, des Rates und der Kommission einen Kompromiss suchen, blieb

das Thema GPAI bis zum Schluss ein Streitpunkt, bis am 9. Dezember 2023 ein Kompromiss gefunden werden konnte. Dieser Verlauf erklärt die eigene und auffallend knappe Regelung von GPAI im V. Kapitel (→ 39 ff.).

Am **21. Mai 2024** hiess der Rat das Verhandlungsergebnis gut. Der AI Act wurde am **12. Juli 2024** im Amtsblatt der Europäischen Union veröffentlicht (ABl. L, 2024/1689, https://dtn.re/0OYJXY).

5 Wann wird der AIA wirksam?

Der AIA ist am 1. August 2024, 20 Tage nach seiner Publikation im Amtsblatt, in Kraft getreten. Seine Bestimmungen werden schrittweise wirksam (Art. 113):

- **2. Februar 2025**: Kapitel I und II (allgemeine Bestimmungen und verbotene Praktiken) werden wirksam.
- **2. August 2025**: Bestimmte Anforderungen einschliesslich Meldepflichten und Sanktionen, werden wirksam. Das betrifft die Bestimmungen zu den notifizierende Behörden und notifizierten Stellen (Kapitel III Abschnitt 4), die Anforderungen an GPAIM (Kapitel V), die Governance in der EU (Kapitel VII) und die Sanktionen (Kapitel XII) sowie die Bestimmungen zur Vertraulichkeitspflicht der Behörden (Art. 78);
- **2. August 2026**: Die meisten Bestimmungen werden wirksam, insbesondere jene für HRAIS, mit der folgenden Ausnahme;
- **2. August 2027**: Die Bestimmungen für HRAIS werden auch im Geltungsbereich von Art. 6 Abs. 1 wirksam, d.h. für AIS, die als Sicherheitsbauteil eines Produkts nach Anhang I verbaut oder als solches verwendet werden.

6 Gibt es im AIA übergangsrechtliche Bestimmungen?

Ja, wenige, nach Art. 111:

- Grundsätzlich gilt der AIA für Betreiber von HRAIS erst ab dem 2. August 2030, wenn die **HRAIS vor dem 2. August 2026 in Verkehr gebracht oder in Betrieb genommen** wurden. Vorbehalten sind aber eine spätere wesentliche Veränderung.
- **Anbieter von GPAIM** fallen erst ab dem 2. August 2027 unter den AIA, wenn das GPAIM vor dem 2. August 2025 in Verkehr gebracht wurde.
- Art. 111 sieht vor, dass AIS, die als Komponenten in **IT-Grossysteme** im öffentlichen Bereich nach Anhang X verbaut werden, erst bis Ende 2030 konform sein müssen. Dabei geht es um das Schengener oder das Visa-Informationssystem und ähnliche Systeme.

7 Was ist “künstliche Intelligenz” (KI, AI)?

Der Begriff “künstliche Intelligenz” (“KI”; “Artificial Intelligence”, “AI”) meint ein Verhalten eines Computers, das zwar nicht intelligent ist und nicht sein kann, von aussen aber wie Intelligenz wirkt. In diese Richtung geht auch eine Definition des Europäischen Parlaments: “Künstliche Intelligenz ist die Fähigkeit einer Maschine, menschliche Fähigkeiten wie logisches Denken, Lernen, Planen und Kreativität zu imitieren”. Bspw. fragt der bekannte Turing-Test dadurch, dass ein Mensch nicht mehr erkennen kann, ob sein Gesprächspartner Mensch oder Maschine ist.

Die Unterscheidung zwischen künstlicher Intelligenz und determinierten Systemen ist deshalb nicht qualitativ, sondern letztlich quantitativ. Künstliche Intelligenz ist, was danach aussieht, weil eine Maschine zu einem Ergebnis gelangt, das nicht von einem Menschen determiniert wurde – oder so wirkt: Determiniert sind auch

komplexe Systeme – sie wirken nur intelligent, weil ihr Ergebnis überraschend ist, was daran liegt, dass eine maschinelle Entscheidung aufgrund der besonderen Komplexität und des fehlenden Zugangs zu den Trainingsdaten *faktisch* nicht in allem nachvollziehbar ist. Dies macht eine Auslegung auch des Begriffs des KI-Modells nach dem AIA schwierig (→ 13).

8 Welche Begriffe definiert der AIA?

Der AIA definiert in Art. 2 ganze 68 Begriffe. Sie werden im AIA anschliessend verwendet, ohne dass jeweils im entsprechenden Artikel ausdrücklich auf die Definition zurückverwiesen wird – man muss bei der Lektüre deshalb öfters zu Art. 2 zurückkehren, zumal auch Begriffe legaldefiniert werden, bei denen man das nicht unbedingt erwarten würde (bspw. “Risiko” oder “weitverbreiteter Verstoss”).

Erschwerend kommt hinzu, dass Begriffe in der Praxis teilweise eher auf Deutsch (“Inverkehrbringen”) und teilweise eher auf Englisch verwendet (“Provider”, “Deployer”). Eine Gegenüberstellung der deutsch- und englischsprachigen Entsprechungen findet sich deshalb im Anhang dieser FAQ.

9 Welche Rolle spielt die Statistik im Bereich der KI?

Beziehungen zwischen Daten werden durch statistische Modelle abgebildet. Das heisst nicht, dass die Statistik für sich genommen eine Form von KI darstellt. Statistische Methoden sind mathematische Modelle, die sowohl bei KI wie auch bei deterministischen Ansätzen eingesetzt werden. Machine Learning (→ 10) und andere Ansätze arbeiten in der Regel aber mit statistischen Methoden.

Ein wichtige solche Methode ist bspw. die **Regressionsanalyse**. Sie bestimmt diejenigen Faktoren (Variablen), die für ein Ergebnis massgebend sind (bzw. die Stärke des Einflusses einer Variabel auf ein Ergebnis), was eine entsprechende Prognose erlaubt. Wenn auf einem Diagramm die x-Achse der Anzahl Besucher bei einer Ausstellung und die y-Achse der Regenfall ist, bezeichnen die Punkte auf dem Diagramm das Besucheraufkommen abhängig vom Regen. Zieht man eine Linie, die mathematisch betrachtet am besten zu allen Punkten passt (“Regressionslinie”), erklärt sie das Verhältnis der Achsen bzw. der Variablen, hier also, wie sich der Regen auf das Besucheraufkommen beeinflusst. Sie kann auch angeben, wie hoch die Abweichung der Datenpunkte vom Soll ist, d.h. den Fehlerbereich der Linie, die Schwankungsbreite, den Verlässlichkeitsgrad der Regressionslinie (i.d.R. mit “R^2” oder “r^2” dargestellt; ein R^2-Wert von 0.73 meint, dass 73% der Daten durch die Regressionslinie erklärt werden).

Eine **lineare Regression** basiert auf der Hypothese, dass der Zielwert (das Besucheraufkommen) linear von einer Variable (dem Regen) abhängt, oder dass der Marktwert einer Immobilie jeweils gleich auf eine Veränderung von Grundstückfläche und Standort reagiert. Hier wird eine gerade Linie durch die Datenpunkte gezogen, und auf dieser Basis können weitere Werte (das Besucheraufkommen, der Immobilienwert) bestimmt werden. Dadurch sind einfache prognostische Modelle möglich. Bei der **nicht-linearen Regression** wird bspw. eine gekrümmte Linie gebildet, weil eine nicht-lineare Beziehung dargestellt werden soll (bspw. wenn das Besucheraufkommen erst bei Starkregen und nicht schon bei Niesel sinkt, oder wenn Verkaufszahlen bei steigenden Preisen nach einem bestimmten Preis – der Preisschwelle – stärker sinken). Auch hier wird mit einer determinierten Logik gearbeitet.

Andere statistische Methoden sind bspw. **Cluster-Analysen**. Es geht dabei nicht um eine bestimmte lineare oder nicht-lineare Beziehung zwischen Variablen, sondern darum, Beziehungen zwischen Daten über Distanz- oder Ähnlichkeitsmasse zu quantifizieren und Objekte mit

niedrigem Distanzmass in eine gemeinsame Gruppe einzuordnen. Bei zwei- oder mehrdimensionalen Daten ("Datenwolken") haben Cluster einen gemeinsamen Schwerpunkt, und Cluster-Analysen dienen dazu, diese Schwerpunkte zu finden und Daten demjenigen Cluster zuzuordnen, dessen Mittelpunkt am nächsten liegt. Das kann bspw. dazu dienen, mögliche Schuldner bei der Kreditvergabe einem Cluster zuzuordnen und die Kreditkonditionen auf dieser Basis zu vergeben.

Dabei lassen sich **parametrische** von **nicht-parametrischen** Modellen unterscheiden. Bei der nicht-parametrischen Regression wird der Zusammenhang zwischen den Variablen nicht vorgegeben, sondern nach unterschiedlichen Kriterien aus vorhandenen Daten erst abgeleitet, bspw. zur Modellierung von Wirtschaftsdaten, zur Untersuchung von Schadstoffkonzentrationen oder zur Prognose von Aktienkursen. Die parametrische Statistik setzt voraus dagegen voraus, dass die verwendeten Daten einer bestimmten statistischen Verteilung entsprechen, die durch eine feste Anzahl von Parametern charakterisiert wird.

10 Was ist "Machine Learning" (ML)?

Aus einer technischen Warte ist KI das Teilgebiet der Informatik, das sich mit der Entwicklung entsprechender Systeme befasst. Die wichtigste Technologie in diesem Gebiet ist "**Maschinelles Lernen**", "**Machine Learning**" oder "**ML**". Es ist kein Synonym für KI, weil ML im Wesentlichen der Erkennung von Mustern und der Ableitung von Prognosen auf deren Basis dient, während KI versucht, eine Aufgabe zu lösen.

ML soll es einem Computer ermöglichen, auf der Basis von Daten zu "lernen", d.h. aus Daten Erkenntnisse abzuleiten. "Erkenntnis" ist allerdings der falsche Begriff. Die alte Unterscheidung zwischen Deduktion und Induktion ist hier wichtig: Bei **deduktiven Schlüssen** wird eine als wahr gegebene Regel angewendet, und die aus der Regel abgeleiteten Ergebnisse können als so wahr wie die Regel selbst gelten (Regel: Alle Fische können schwimmen; Input: Wanda ist ein Fisch; Ergebnis: Wanda kann schwimmen). Es gibt daneben auch die **Abduktion**. Denn Kopfschmerzen können viele Ursachen haben; der Rückschluss wäre deshalb unzulässig. Abduktion arbeitet also mehrere mögliche Kausalketten ab, um die wahrscheinlichste Ursache zu finden. Solche Systeme sind häufig; ein bekanntes Beispiel ist das Diagnosesystem "CADUCEUS". Bei **induktiven Schlüssen** wird demgegenüber aus Informationen auf eine vermutete Regel geschlossen. Machine Learning geht häufig induktiv vor: Aus Daten werden statistisch begründete Aussagen erzeugt, die mehr oder weniger überzeugende Hypothesen sind, aber keine Wahrheit oder Objektivität für sich in Anspruch nehmen können. Die Grenzen sind allerdings fliessend, weil diese Ansätze auch kombiniert werden können.

Durch ML kann eine Maschine also Daten beobachten und gestützt darauf Prognosen oder Hypothesen erzeugen, die mehr oder weniger wahrscheinlich sind, d.h. die mehr oder weniger gut zu den Eingangsdaten passen. Die Erklärung der so gebildeten Hypothese – bspw. der Schluss einer Korrelation auf Kausalität – steht ausserhalb des ML; sie ist eine Form der Heuristik, nicht des ML. Deshalb ist ML häufig auf grosse Datenmengen angewiesen: Die verstreckten Muster, die Beziehungen zwischen Daten, werden erst in der Masse beobachtbar.

Muster erkennen heisst verallgemeinern. Je besser ein Modell – Modelle sind mathematische Funktionen – verallgemeinern kann, desto leistungsfähiger ist es. Dazu dient, wie erwähnt, ein Training. Verwendet das Training zu wenige Daten, kann es keine zuverlässigen Schlüsse ziehen – man spricht von Unteranpassung oder "**underfitting**". Umgekehrt kann ein Modell die Inputdaten zu gut lernen, im Extremfall lernt es sie auswendig. Dann passt es zu den Inputdaten, kann aber schlecht verallgemeinern, wie ein Mensch, der zwar ein gutes Gedächtnis hat, aber nicht denkt. Diese Überanpassung nennt man "**overfitting**". Für das Training einer ML werden deshalb neben den Trainings- auch Validierungs- und Testdatensätze verwendet, um sowohl ein Over- als auch ein Underfitting zu re-

duzieren und die Aussagekraft – die Verlässlichkeit der verallgemeinernden Hypothese – zu verbessern oder zumindest einzuschätzen.

ML verwendet statistische Modelle (→ 9), bspw. die lineare Regression vor allem beim überwachten Lernen oder Cluster-Analysen beim unüberwachten Lernen. ML kann dabei sowohl parametrische als auch nicht-parametrische Methoden verwenden. Parametrische Modelle bei ML verwenden zwar eine festgelegte Modellstruktur, aber die Werte der Parameter werden durch ein Training optimiert. Ein Beispiel ist lineare Regression, wenn ein Modell lernt, Immobilienpreise zu prognostizieren, weil es die statistischen Zusammenhänge zwischen bestimmten Parametern und den Preisen im Lauf des Trainings verlässlicher zu erkennen lernt. Diese Modelle erfordern also, dass bestimmte Annahmen über die Daten gelten, schliessen ein "Lernen" durch Training aber nicht aus. Nichtparametrische Modelle im ML haben dagegen keine feste Struktur oder keine feste Anzahl von Parametern. Beispiele sind Entscheidungsbäume, die im Lauf eines Trainings verbessert werden. Sie sind ebenfalls statistische Modelle, die an jedem Knoten den bestmöglichen "Split" auf Basis statistischer Kriterien bestimmen.

Ein besseres Kriterium zur Unterscheidung zwischen ML und deterministischen Ansätzen ist deshalb das grundsätzliche Vorgehen: **Deduktive Methoden** verwenden bestimmte Grundannahmen und ziehen daraus Schlüsse, während **induktive Methoden** mögliche Regeln durch einen Trainingsvorgang mit zunehmender Verlässlichkeit generieren. Die Regelgenerierung ist also ein wesentlicher Faktor bei der Unterscheidung zwischen deterministischen und nicht-deterministischen Ansätzen. Ein Beispiel sind Entscheidungsbäume, die nicht vordefiniert, sondern durch ein Training generiert werden – das Modell ermittelt im Training diejenigen Regeln, die die Trainingsdaten am besten trennen oder erklären, d.h. die die grösste Aussagekraft für eine Zielvariable (z.B. Kreditwürdigkeit) haben. Diese Regeln sind interpretier- und weiterverwendbar. Ein weiteres Beispiel sind Assoziationsanalyen, die Beziehungen in grossen Datenmengen darstellen und Regeln generieren, die häufige Zusammenhänge beschreiben. Bei der Warenkorbanalyse kann bspw. eine Regel wie "Wer am Freitagabend Windeln kauft, kauft auch Bier" generiert werden. Auch diese Regeln sind explizit und können interpretiert werden.

Expertensysteme sind Systeme, die eine Wissensbasis enthalten, bspw. anwendungsspezifische Wenn-/Dann-Regeln. Auf diese Wissensbasis wendet das System Regeln an, um weitere Fakten oder Schlussfolgerungen abzuleiten (Inferenz). Das System kann dabei Wahrscheinlichkeiten angeben und u.U. auch mit ungenauen Angaben arbeiten ("fuzzy logic"). Ein Beispiel eines Expertensystems ist das bekannte "Mycin", ein in den 70er-Jahren an der Universität Stanford entwickeltes System, das zur Unterstützung des Antibiotika-Einsatzes entwickelt wurde. Auf der Grundlage von Parametern wie Erregertyp, Krankheitsverlauf und Labordaten konnte das System durch bestimmte Regeln Entscheidungen auf Basis von Wahrscheinlichkeiten und Unsicherheiten treffen oder vorbereiten.

Während Entscheidungsbäume explizite Regeln generieren, sind neuronale Netzwerke ein Beispiel für eine implizite Regelgenerierung. Neuronale Netzwerke lernen komplexe Muster aus den Daten, aber die „Regeln", die sie anwenden, um Vorhersagen zu treffen, sind in den Gewichtungen und Aktivierungen der Neuronen verborgen. Obwohl es in neuronalen Netzen keine expliziten „Wenn-Dann"-Regeln gibt, werden die Entscheidungen dennoch durch Regeln bestimmt, die während des Trainingsprozesses gelernt wurden.

Die Schwierigkeit bei neuronalen Netzen besteht darin, dass die Regeln oft schwer verständlich sind – sie sind „black box"-Modelle. In jüngster Zeit gab es jedoch Fortschritte in der erklärbaren KI ("Explainable AI"), die darauf abzielt, diese impliziten Regeln offenzulegen und verständlicher zu machen.

Wie ML vorgeht, ist damit noch nicht gesagt. Nach der Methodik des Lernens können vier Formen unterschieden werden:

- **Überwachtes Lernen** (supervised learning), bei dem gekennzeichnete Datensätze (Labeling) verwendet werden.
- **Unüberwachtes Lernen** (unsupervised learning), bei dem Muster ohne Labeling erkannt werden, beispielsweise beim Data Mining.

- **Semi-überwachtes Lernen** (semi-supervised learning) als Zwischenform, bei der sowohl gelabelte als auch ungelabelte Daten verwendet werden.
- **Bestärkendes Lernen** (re-inforcement learning), bei dem das Lernen durch Interaktion mit der Umgebung verstärkt wird.

Hilfreich ist weiter die Unterscheidung zwischen **symbolischem** und **subsymbolischem Lernen**. Symbolisches Lernen heisst so, weil es Symbole und logische Regeln zur Repräsentation von Wissen verwendet. Ein Beispiel sind **Entscheidungsbäume** ("decision trees"): Hier wird eine Struktur von Bedingungen oder Regeln analog zu einem Flussdiagramm verwendet oder generiert, um aus Trainingsdaten regelbasiert Schlüsse zu ziehen. Die Struktur ist baumartig, weil Knoten Entscheidungen darstellen – jeder Knoten entspricht einer Wenn/Dann-Regel, basierend auf einer Eigenschaft der Eingabedaten. Die Zweige stellen die Ergebnisse der Anwendung dieser Regeln dar, und die Blätter stehen als Endpunkte für das Ergebnis, die Klassifizierung oder Vorhersage. Entscheidungsbäume arbeiten also durch definierte Entscheidungsprozesse. Mehrere Entscheidungsbäume können dabei auf unterschiedlichen Daten trainiert werden und dann zusammengenommen bspw. durch Mehrheitsentscheid bessere Ergebnisse liefern als ein einzelner Baum, der zu Overfitting neigt.

Symbolisches Lernen kann bei grossen Datenmengen allerdings an Grenzen stossen. **Subsymbolisches Lernen** verwendet dagegen Rohdaten, die nicht in systemkompatible Symbole umgewandelt werden müssen. Dieses Vorgehen eignet sich besser für die Erkennung komplexer Muster in Inputdaten, ist u.U. aber weniger transparent, weil die komplexen Prozesse schwerer nachzuvollziehen sind. Wann welche Form von ML zur Anwendung kommt, ist nicht vom Einsatzgebiet abhängig, sondern eher davon, ob Regeln bereits bekannt sind oder erst gebildet werden sollen. Im Beispiel der Bonitätsbewertung kann ein Unternehmen nicht nur mit einem Entscheidungsbaum arbeiten; es kann auch versuchen, Korrelationen zwischen Verlusten und anderen Faktoren wie bspw. Alter, Wohnort, Geschlecht, Einkaufsverhalten, Haushaltsgrösse etc. durch eine Form subsymbolischen ML erst festzustellen. Im Anschluss können die beobachteten Zusammenhänge als Regeln eines Entscheidungsbaums verwendet werden.

Zum subsymbolischen Lernen gehören bspw. **künstliche neuronale Netze** (und Deep Learning als Schlagwort für besonders komplexe Netze → 11).

11 Was sind neuronale Netze?

Neuronale Netze sind Algorithmen, die die Informationsverarbeitung im Gehirn nachbilden, um Muster in Inputdaten zu erkennen. Dabei wird eine grosse Zahl verbundener "Knoten" verwendet, die zusammen "Schichten" bilden und die Eingabedaten schrittweise u.U. über mehrere oder sehr viele Schichten hinweg verarbeiten ("gewichten"). Im Unterschied zu Entscheidungsbäumen (→ 10) sind neuronale Netze komplexer verbunden, weil jeder Knoten mit mehreren anderen Knoten der nächsten Schicht verbunden sein kann.

Damit das Netz zu einer sinnvollen Verarbeitung in der Lage ist, müssen diese Gewichtungen richtig eingestellt werden. Entsprechend erfolgt die Entscheidungsfindung in neuronalen Netzen als verteilte und kontinuierliche Verarbeitung von einem aufnehmenden "**Input Layer**" über dazwischengeschaltete "**Hidden Layer**" zur Ausgabeebene, dem "**Output Layer**"; wobei das Netz durch Anpassung der Gewichte zwischen den Knoten lernt. Entscheidungsbäume arbeiten dagegen mit expliziten Bedingungen ("wenn A > X gehe nach links, sonst rechts"). Jeder Knoten trifft eine Entscheidung, die zu einem bestimmten Zweig führt, weshalb die Entscheidungswege grundsätzlich vollständig nachvollziehbar sind. Deduktive Systeme sind deshalb eher eine "White Box", induktive Systeme eine "Black Box".

Die Faktoren der angesprochenen Gewichtung der Knoten im Netz wird durch Training verbessert, indem der Output des Netzes mit einem er-

warteten Ergebnis verglichen wird. Bei Abweichungen werden die Gewichte durch weitere Trainingsdaten angepasst – und so weiter.

Dabei kann dieses Training wiederum unterschiedlich erfolgen:

- Beim **überwachten Lernen** werden dem Netz sowohl Inputdaten als auch die gewünschten Ausgaben zur Verfügung gestellt. Das Netz lernt so, die Beziehung zwischen Inputdaten und Output abzubilden. Ein Beispiel ist ein Input von Tierfotos, wenn dem Netz gleichzeitig ein Datensatz mit entsprechend gekennzeichneten Bildern von Hunden und Katzen ("Labels") bereitgestellt wird. Das Netz vergleicht die Prognosen aus den Inputdaten mit den Labels und passt die Gewichtungen solange an, bis Prognosefehler minimiert werden. Das ist ein übliches Vorgehen für Datenklassifikationen, bspw. bei der Bildklassifikation, bei Spamfiltern (Lernen durch Markierung von E-Mails als Spam) oder der Prognose von Immobilienpreisen (= die gelabelten Daten) auf der Grundlage von Angaben über Grösse, Lage und Ausstattung der Immobilie.

- Beim **unüberwachten Lernen** erhält das Netz Inputdaten, aber keine Labels. Muster und Strukturen in den Daten muss es also selbständig erkennen, indem es ähnliche Datenpunkte gruppiert oder Daten auf bestimmte relevante Merkmale reduziert. Dieses Vorgehen eignet sich für Datenexplorationen, bspw. bei der Kundensegmentierung (Gruppierung anhand des Kaufverhaltens ohne vorgegebene Kategorien), der Erkennung ungewöhnlicher Transaktionen ohne Definition von "ungewöhnlich" oder der Erkennung von Themenclustern in einer grossen Textsammlung.

- **Semi-überwachtes Lernen** kombiniert überwachtes und unüberwachtes Lernen – für das Training werden sowohl (wenige) gelabelte als auch (sehr viele) ungelabelte Daten verwendet. Durch die Labels werden Muster eher erkannt. Wenn das Labeln von Daten zu aufwendig ist, kann dieses Vorgehen sinnvoll sein, etwa wenn eine kleinere Zahl gelabelter Röntgenbilder mit einer grösseren Menge nicht klassifizierten Bilder zur Verbesserung der Diagnosegenauigkeit verwendet wird, wenn klassifizierte Produktbewertungen mit unklassifizierten Bewertungen verwendet wird, um die Stimmung in neuen Bewertungen zu bestimmen ("Sentimentanalyse"), oder bei der Spracherkennung, wenn transkribierter Audioaufnahmen mit Sprachdaten zur Verbesserung der Erkennungsgenauigkeit kombiniert werden.

- Beim **verstärkenden Lernen** ("Reinforcement Learning") interagiert das Netz mit einer Umgebung und "lernt" – passt Gewichtungen an – durch Belohnung und Bestrafung. Es ist ein interaktives Trial- and Error-Vorgehen, das bspw. für das Training eines Agenten bei Spielen wie Schach oder Go verwendet wird (Lernen durch wiederholtes Spielen, bei der Roboternavigation (Lernen durch Navigation in einer Umgebung) oder beim Energiemanagement (Lernen durch Anpassung der Stromverteilung basierend auf Verbrauchsmustern).

Wie andere ML-Modelle bilden auch neuronale Netze Regeln. Allerdings sind diese Regeln nicht explizit, anders als bspw. bei einer Assoziationsanalyse (→ 9). Sie wollen es auch nicht sein – das Ziel besteht nicht in der Regelfindung, sondern in einem Output, das Regeln anwendet, aber nicht darstellt ("Black Box"). Ein Entscheidungsbaum als bspw. bildet also eine **explizite Regel**, während neuronale Netze **implizite Regeln** generieren – diese Regeln sind in Aktivierungen und Gewichtungen der "Neuronen" versteckt. Das Problem bei neuronalen Netzen besteht also darin, dass die Regeln oft schwer verständlich sind.

Es gibt allerdings Ansätze, implizite Regeln offenzulegen. Bspw. visualisieren "Saliency Maps", welche Bestandteile des Inputs am meisten zur Entscheidung beigetragen haben (bspw. durch Hervorhebung des Bildbereichs, der für die Klassifizierung ausschlaggebend war), und ähnlich funktionieren "Local Interpretable Model-agnostic Explanations" (LIME) – sie verwenden einfache Modelle wie bspw. lineare Regressionen parallel zum Einsatz des neuronalen Netzes und können nachvollziehbare Erklärungen liefern (z.B. dass Wörter wie "gratis" für die Klassifizierung einer E-Mail als Spam massgebend sind).

12 Was ist ein Large Language Model (LLM)?

Ein Large Language Model (**LLM**) basiert auf einem neuronalen Netz (→ 11) und "versteht" Sprache. Bekannte Beispiele sind die GPT-Modelle von OpenAI, Gemini von Google, LLaMA von Meta, Claude von Anthropic, Command von Cohere, Grok von X, die Modelle von Mistral, Ernie von Baidu oder Falcon des Technology Innovation Institute in Abu Dhabi.

Beim **Training eines LLM** können eine vorangehende Aufbereitung und das eigentliche Training unterschieden werden.

Im Rahmen des **Preprocessing** werden die Trainingsdaten (bspw. Texte aus Büchern, Websites, Foren, Wikipedia etc., inzwischen auch auf Basis entsprechender Lizenzen grosser Verlage wie der NY Times; zum Training → 36) bereinigt. Bspw. werden irrelevante oder fehlerhafte Inhalte oder Spam entfernt, teils überflüssige Symbole und Stoppwörter wie "der", "die" etc.).

Ein "Tokenizer" zerlegt Texte anschliessend in kleinere Einheiten (die **Tokens**), je nachdem ein Wort, ein einzelnes Zeichen oder ein Wortbestandteil. Letzteres trifft bspw. bei OpenAI zu, hier kommt eine Variante des "Byte-Pair Encoding" zum Einsatz, bei dem von Einzelzeichen ausgehend die häufigsten Zeichenpaare zu neuen Tokens zusammengefügt werden, wodurch das Vokabular sukzessive wächst und häufigere Wörter oder Bestandteile werden als Ganzes verwendet werden. Homonyme wie "Bank" können dabei je nach Kontext als mehrere Tokens gespeichert werden ("Geld auf der Bank", "auf der Bank sitzen").

Die Tokens haben für sich genommen aber keine Aussagekraft – interessant sind sie erst in **ihrer Beziehung zu anderen Tokens**. Diese Beziehungen ergeben sich im Lauf des Trainings aus den Inputdaten und können konzeptionell als Nähe- bzw. Distanzwert ausgedrückt werden. Bspw. hat das Wort "Haus" eine grössere Nähe zum Wort "Dach" als das Wort "Schaden", das Token "gross" hat eine grössere Nähe zu "artig", usw. Jedem Token werden deshalb entsprechende Werte zugeordnet. Diese Werte sind die **Vektoren**: Allgemein ist ein Vektor eine geordnete Liste von Zahlen, die mit einer bestimmten Dimensionalität in einer bestimmten Reihenfolge angeordnet sind. Im Kontext eines LLM ist ein Vektor der Wert eines Tokens als Beziehung zu anderen Tokens. Die gelernten Vektoren werden dabei als "**Embedding**" gespeichert – Embeddings sind also Ausdruck der Struktur oder der Eigenschaften von Daten.

"**Dimensionalität**" meint dabei die Anzahl Zahlenwerte des Vektors. Diese Zahlen drücken dabei die Eigenschaften eines Tokens aus. Ein Vektor mit einer Dimensionalität von 768 meint also eine Reihe von 768 Zahlen, wobei jede für ein bestimmtes gelerntes Merkmal steht. Je höher die Dimensionalität, desto feiner die erfassten Bedeutungsunterschiede. Das Modell GPT-3 von OpenAI hat eine Dimensionalität von 768 bis 12'288, je nach Variante. Bei GPT-4 ist der Wert nicht bekannt, vermutlich aber ähnlich. Jedes Token erhält im Training also bis zu 12'288 Eigenschaften.

Austrainierte Modelle können anschliessend für bestimmte Anwendungsgebiete auf einem spezifischen, kleineren Datensatz weiter trainiert werden ("**Finetuning**"), bspw. durch medizinische Daten, eine technische Dokumentationen, juristische Texte oder Material eines bestimmten Unternehmens. Das Modell wird auf diesen Daten so weitertrainiert, dass es die gelernten Fähigkeiten verfeinert, ohne sie zu verlernen. Dabei werden die Parameter des Modells Leicht angepasst – das Modell lernt bspw. Fachbegriffe, bestimmte Formulierungen oder typische Satzstrukturen. Ein Beispiel ist der EDÖBot von datenrecht (https://edoebot.datenrecht.ch), der auf einem Modell von OpenAI basiert, aber mit datenschutzrechtlichen Material weiter trainiert wurde.

Die Leistungsfähigkeit kann auch durch "**Retrieval-Augmented Generation**" ("RAG") verbessert werden. Hier wird ein LLM mit externen Informationsquellen kombiniert, d.h. es werden Informationen ausserhalb des Modells bei der Abfrage einbezogen, bspw. aktuellere oder spezifischere Informationen, die im Training nicht gelernt wurden. Eine Suchkomponente ("**Retriever**") durchsucht bei der Abfrage eine externe Datenbank nach relevanten Daten, der **Generator** verwendet diese Daten für eine bessere Ant-

wort. Auch dies kommt beim EDÖBot zur Anwendung, er kann bspw. auf die Botschaft zum aktuellen DSG oder die Leitfäden des EDÖB zugreifen.

Grundfragen

13 Was ist ein “KI-System” (AI System, AIS)?

Im Verlauf der Verhandlungen (→ 4) war dieser zentrale Punkt – der über die sachliche Anwendbarkeit des AIA bestimmt – einer der besonders strittigen Punkte, und man kann nicht behaupten, das Ergebnis sei geglückt. Der Kommissionsentwurf vom April 2021 (https://dtn.re/1Lpgyz) hatte ein KI-System noch über in einem Anhang definierte Techniken und Konzepte definiert und neben ML-Ansätzen auch logik- und wissensgestützte Konzepte und “statistische Ansätze, Bayessche Schätz-, Such- und Optimierungsmethoden” erfasst. Im Trilog einigte man sich dann auf die heutige, bewusst an die Definition der OECD (https://dtn.re/dzZqxl)angelehnte Begriffsbestimmung.

Der AIA definiert ein AIS nun wie folgt (Art. 3 Nr. 1 und ErwG 12):

> *„KI-System” 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;*

Es geht also um

- “ein maschinengestütztes System” (also kein bspw. biologisches System – das Verpflanzen eines Gehirns wäre also kein Inverkehrbringen eines AIS),
- 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”.

Massgebend sind im Ergebnis zwei Elemente, die aber letztlich in eines zusammenfallen:

- Das System ist für einen **autonomen Betrieb** ausgelegt. Das heisst nach ErwG 12, dass es nicht “ausschließlich von natürlichen Personen definierten Regeln für das automatische Ausführen von Operationen beruht”, sondern dass es “bis zu einem gewissen Grad unabhängig von menschlichem Zutun agieren und in der Lage [ist], ohne menschliches Eingreifen zu arbeiten”; und
- es kann aus Input einen Output ableiten, wobei es dabei nicht um irgendeine Ableitung geht, sondern um “Lern-, Schlussfolgerungs- und Modellierungsprozesse” (ErwG 12; “Inferenz”).

Das lässt allerdings die Frage offen, was mit der erforderlichen Autonomie im Betrieb gemeint ist.

Die Basis für ein AIS wird vor allem **Machine Learning** (ML; Q10) sein (ErwG 12: “maschinelles Lernen, wobei aus Daten gelernt wird, wie bestimmte Ziele erreicht werden können”). Es läge an sich nahe, hier die erwähnte Unterscheidung zwischen deduktiven und induktiven Modellen zu verwenden (→ 10) und AIS als ML zu verstehen, die im Unterschied zu deterministischen statistischen Modellen nicht deduktiv vorgeht, d.h. die nicht oder nicht allein vordefinierte Regeln anwendet, sondern Regeln definiert oder vordefinierte Parameter zumindest gewichten lernt. Ein AIS wäre demnach bspw. ein Modell, das aus Trainingsdaten lernt, wie stark sich die Grundstücksfläche als vorgegebener Parameter auf Immobilienpreise auswirkt, und kein AIS wäre ein Modell, das definierte Parameter und Gewichtungen auf neue Daten anwendet – bspw. ein einfaches Excel mit einer entsprechenden Formel.

So klar ist die Unterscheidung allerdings nicht. Nach ErwGr 12 erfasst der AIA als AIS auch "logik- und wissensgestützte Konzepte, wobei aus kodierten Informationen oder symbolischen Darstellungen der zu lösenden Aufgabe abgeleitet wird". Das trifft auf das erwähnte Beispiel zu: Das Excel zur Berechnung der Immobilienpreise ist ein logik- und wissensgestütztes Konzept, das aus der kodierten Aufgabe (die Excel-Formel) Ableitungen trifft (die Immobilienpreise abhängig von den Inputdaten berechnet). Ob dieses Konzept, also die Excel-Formel, auf einem Training beruht, ist an sich belanglos, weil das Excel im Einsatz nicht lernt. Würde man nur von der Unterscheidung zwischen deduktivem und induktivem Vorgehen ausgehen, fielen alle diese Systeme aus der Definition.

Es kann jedenfalls nicht darum gehen, dass sich das Modell **im laufenden Betrieb**, nach der Inbetriebnahme, weiter verändert, aus zwei Gründen: Zum einen ist das Element der Anpassungsfähigkeit nach dem Wortlaut der Bestimmung nicht zwingend, sondern illustrativ. Zum anderen wären austrainierte Modelle sonst nicht vom AIA erfasst, und das betrifft die grosse Mehrheit der verwendeten Systeme, auch verbreiteten LLMs, was natürlich nicht beabsichtigt ist. Ein austrainiertes System ist im Betrieb nun aber nicht wirklich autonom – es verarbeitet die Inputdaten seinen Parametern entsprechend, die zwar in einer Trainingsphase erlernt sein mögen, sich wie erwähnt aber nicht mehr verändern (bis zu einem Update und vorbehaltlich des Ausnahmefalls, dass ein System im Betrieb weiter trainiert, wie es bspw. bei Systemen zur Betrugsbekämpfung der Fall sein kann). So betrachtet sind die meisten Systeme deterministisch, nicht autonom.

Man kann auch nicht alleine auf die **Entwicklungsphase** schauen. In der Entwicklung kann ein Modell zwar lernen, und weil das Ziel des Lernvorgangs nur funktional umschrieben wird (bspw. verlässliche Klassifizierung von Bildern, Generierung eines sinnvoll wirkenden Textes), nicht aber technisch, ist der Lernvorgang auf einer technischen Ebene nicht determiniert (wie die Parameter einzustellen sind, damit das Lernziel erreicht wird, ist nicht vorgegeben, deshalb das Training). Der Wortlaut von Art. 3 Nr. 1 nimmt aber ausdrücklich auf den "Betrieb" und nicht das Training Bezug – das Training wird im AI Act angesprochen (→ 36), aber nicht in der Definition des AI-Systems, und es ist anders als das Testen auch nicht zwingend vorgeschrieben. Man kann die erforderliche Autonomie also nicht nur im Training suchen.

Damit bleibt weiterhin offen, was gemeint ist. Die **OECD** hat für ihre parallele Definition des "AI System" im März 2024 aber ein Begleitmemorandum veröffentlicht, das sich zur erforderlichen Autonomie etwas klarer äussert:

> *AI system autonomy (contained in both the original and the revised definition of an AI system) means the degree to which a system can learn or act without human involvement following the delegation of autonomy and process automation by humans. Human supervision can occur at any stage of the AI system lifecycle, such as during AI system design, data collection and processing, development, verification, validation, deployment, or operation and monitoring. Some AI systems can generate outputs without these outputs being explicitly described in the AI system's objective and without specific instructions from a human.*

Autonomie im Betrieb bezieht sich also nicht auf die Funktion des Systems als solche, die wie angemerkt i.d.R. determiniert ist, sondern auf das, was es mit Inputdaten tut: Ein System ist autonom, wenn es nach dem Input ohne menschliches Eingreifen arbeiten kann und dabei einen Output generiert, der nicht explizit vorgegeben ist. Das Nicht-deterministische ist also in der Datenverarbeitung zu suchen und bezieht sich auf **das Verhältnis von Input und Output**.

Man kann dem zwar entgegenhalten, dass es echte Autonomie auch hier nicht gibt. Wenn das System austrainiert ist, ist seine Datenverarbeitung durch die Parameter des Systems determiniert. Der gleiche Input muss den gleichen Output erzeugen, es sei denn, es sei eine Zufallsfunktion eingebaut. Das ist zwar häufig der Fall, bspw. bei den Modellen von OpenAI (mit der Temperatureinstellung kann dieser Faktor bis zu einem gewissen Grad gesteuert werden), aber auch ein Zufallsgenerator ist im Grunde determiniert (nicht-deterministische Generatoren liefern bei gleichen Ausgangsbedingungen unterschiedliche Werte, aber weil die Software für sich genommen deterministisch ist, muss zur Randomisierung ein externer Faktor wie bspw.

radioaktiver Zerfall einbezogen werden, und dieser Faktor gehorcht Naturgesetzen).

Der AIA ist aber ein Gesetz mit einem Zweck und keine naturphilosophische Betrachtung. Entsprechend muss man ihn funktional auslegen, insbesondere mit Blick auf die Rechtsfolgen, die Sachverhalte erfassen sollten, auf die sie ausgelegt sind. Als Zwischenergebnis muss man die erforderliche Autonomie also auf der Ebene der **Datenverarbeitung** erfolgen, im Weg vom Input zum Output, und dieser muss so beschaffen sein, dass das Ergebnis in einer normalen menschlichen Betrachtung **nicht determiniert erscheint**.

Letztlich ist das eine Form des Turing-Tests (→ 6): Der AIA erfasst ein System dann als AI-System, wenn es wie AI aussieht. In diese Richtung geht auch die Faustregel der österreichischen Datenschutzbehörde (→ 1):

> *Vereinfacht gesagt handelt es sich um Computersysteme, die Aufgaben ausführen können, **die normalerweise menschliche Intelligenz erfordern**. Das bedeutet, dass diese Systeme Probleme lösen, lernen, Entscheidungen treffen und mit ihrer Umgebung interagieren können, ähnlich wie Menschen es tun.*

Ein AIS ist also ein System, das im Betrieb, bei der Generierung eines Outputs, aus verschiedenen, *a priori* gegebenen Möglichkeiten auswählen kann, ohne dass die Auswahl rein zufällig erfolgt und ohne dass sie einer direkten menschlichen Anleitung folgt, **und das daher eine Aufgabe erfüllt, bei der ein Mensch denken müsste**. Dies erklärt auch die Abgrenzung zum determinierten System: Ein Mensch, dem im Detail vorgegeben wird, wie er vorzugehen hat, muss nicht mehr denken. Es wird entsprechend nicht möglich sein, bei allen Systemen trennscharf zu sagen, ob sie unter den AIA fallen oder nicht.

AIS sind etwa:

- Chatbots
- Empfehlungssysteme bei Streaming-Diensten
- Sprachassistenten, die durch Nutzerinteraktion lernen
- autonome Fahrzeuge, die ihre Fahrweise durch Sensor- und Umgebungsdaten anpassen
- Gesichtserkennungssysteme, deren Genauigkeit durch die Verwendung verbessert wird
- ML-basierte Übersetzungstools
- Betrugserkennungssysteme bei Banken, die verdächtige Muster erkennen lernen
- diagnostische Systeme im Gesundheitsbereich
- personalisierte Lernplattformen (schon dann, wenn sie gestützt auf den Lernerfolg Wiederholungsintervalle generieren)
- Spamfilter

Vorausgesetzt bleibt jeweils ein nicht-deterministisches Vorgehen. Reine wenn/dann-Logiken genügen aber nicht, bspw. ein Musikstreamingdienst allen Hörern von Metallica prinzipiell Megadeth vorschlägt – solche Logiken sind deterministisch, eine Lern- oder Schlussfolgerungskomponente fehlt (vorausgesetzt, dass das System nicht selbst auf diese Korrelation gestossen ist).

Keine AIS sind bspw.:

- Excelkalkulationen, allerdings mit dem Vorbehalt, dass auch ein Excel-Dokument zu einem AIS programmiert werden könnte
- Datenbanken wie MySQL, die Informationen auf Abfrage liefern
- Bildbearbeitungssoftware, soweit sie deterministisch ist, d.h. keine Bilder generiert und auch nicht auf einem LLM basiert
- Mailclients, die E-Mails nach fixen Regeln in Ordner verschieben
- der Browser, mit dem ChatGPT verwendet wird
- Spamfilter basierend alleine auf White-/Black-Listen
- eine deterministische Software, die durch oder mit Hilfe von AIS erzeugt wurde (das dürfte heute einen grossen Teil der Software

betreffen, wenn die Entwicklung AI-unterstützt erfolgt, bspw. beim Einsatz von Github Copilot)

Diverse weitere Anwendungsbeispiele finden sich übrigens im Atlas von Algorithm Watch (https://dtn.re/ggJqKy).

Ebenfalls kein AIS ist ein AI-Modell, also Grundlagentechnologie, die noch keinem Anwendungsbereich zugeführt wurde (→ 39).

AIS kann dabei **selbst ein Produkt** sein (bspw. ein AIS zur Einschätzung der Eignung von Stellenbewerbenden), oder es kann als "**embedded system**" oder "embedded AI" Teil eines anderen Produkts sein (z.B. ein Steuerungssystem). Bei Steuerungssystemen wird das entsprechende Produkt also nicht insgesamt zum AIS, wie sich aus Art. 6 Abs. 1 und Art. 25 Abs. 3 ergibt – es unterliegt weiterhin den entsprechenden Produktevorschriften, aber das "embedded AIS" wird durch den Einbau zum HRAIS, sofern das Produkt unter Anhang 1 fällt (→ 28). Erst wenn der Produktehersteller die AIS-Komponente im eigenen Namen auf dem Markt bereitstellt oder in Betrieb nimmt, wird er zum Anbieter des HRAIS (Art. 25 Abs. 3). Bei der Konformitätsbewertung wird das Steuerungssystem dennoch im Kontext des Gesamtsystems zu bewerten sein. Bei anderen AIS ist eine Aufteilung dagegen nur möglich, wenn die AI-Komponente von anderen Komponenten klar abgrenzbar ist (bspw. in einem Recruiting-System, das ein KI-Modul für das Bewerberranking klar von der Verwaltung der berücksichtigten Bewerbungen trennt).

Mit der Qualifikation als AIS als solcher ist nichts über das damit verbundene Risiko gesagt – auch deshalb nicht, weil sich die Risiken nicht aus der Technologie, sondern den Bedingungen ihres Einsatzes ergeben. Der AI Act unterteilt AIS-Use Cases dabei grundsätzlich – wenn auch nicht expressis verbis – in vier Kategorien (→ 16). Daneben kennt der AIS die GPAI, deren Regelung bei der Verhandlung ein pièce de résistance war (→ 39 ff.).

14 Fallen alle AI-Systeme unter den AIA?

Nein. Zunächst kann die EU nur innerhalb ihres Mandats regulieren, d.h. nur Tätigkeit im Anwendungsbereich des Unionsrechts. Das schliesst Tätigkeiten der Mitgliedstaaten aus, die die nationale Sicherheit betreffen. Bestimmte KI-Systeme sind sodann vom AIA ausgenommen (Art. 2):

- AIS, die ausschliesslich **militärischen Zwecken** und der **nationalen Sicherheit** dient (Art. 2 Abs. 3), womit der AIA die Grenze des EU-Rechts aufnimmt;
- AIS, die ausschliesslich für **Forschungszwecke** entwickelt und verwendet wird (Art. 2 Abs. 6), damit die Forschungsfreiheit nicht beeinträchtigt wird (AIS, deren Einsatzmöglichkeiten die Forschung lediglich umfassen, fallen aber unter den AIA; ErwG 23);
- AIS, die **Privatpersonen** für nicht-gewerbliche Zwecke nutzen (Art. 2 Abs. 10; bspw. der private Einsatz von ChatGPT für die Planung einer Hochzeitsfeier);
- **FOSS** (Art. 2 Abs. 12), d.h. freie und quelloffene Software (bzw. Modelle), unter der Voraussetzung, dass die offene Weitergabe erlaubt ist und Nutzer das Modell kostenlos verwenden, verändern und weitervertreiben dürfen, und mit dem Vorbehalt, dass FOSS erfasst bleibt, wenn es sich um ein HRAIS handelt (→ 28), wenn sie bzw. ihre Nutzung eine verbotene Praxis darstellt (→ 27) oder wenn sie direkt mit Nutzern interagiert oder für die Generierung von Content verwendet wird (Art. 50 → 37);
- AIS während der **Forschungs-, Test- und Entwicklungsphase** vor dem Inverkehrbringen oder der Inbetriebnahme, ausser bei Tests unter Realbedingungen (Art. 2 Abs. 8). Anbieter von AIS müssen aber selbstverständlich auch während dieser Phasen die Anforderungen einhalten oder vielmehr ihre Einhaltung vorbereiten.

15 Was ist allgemein der Regelungsansatz des AIA?

Der AIA ist trotz seines Namens weder eine umfassende Regelung der künstlichen Intelligenz noch Marktverhaltensrecht, sondern Produktesicherheitsrecht. Er orientiert sich an den etablierten Prinzipien der Produktregulierung im europäischen Binnenmarkt, insbesondere in den "New Approach"-Regelungen.

Der "**New Approach**" (oder "Neues Konzept"; siehe dazu die Mitteilung der Kommission KOM(2003)0240 von 2003, https://dtn.re/0mGegd) ist ein Konzept, das die EU in den 1980er Jahren für die Regulierung des Binnenmarktes eingeführt hat: Statt detaillierte technische Vorschriften zu erlassen, legt die EU grundlegende Anforderungen für Produkte als Voraussetzung des Marktzugangs fest. Detaillierte Anforderungen werden dann durch europäische Normungsorganisationen (z.B. CEN, CENELEC oder ETSI) entwickelt. Diese Normen sind nicht zwingend, aber ihre Einhaltung begründet die Vermutung der Konformität der entsprechenden Produkte (im AIA: Art. 40).

Der Nachweis der Konformität erfolgt sodann im **Konformitätsbewertungsverfahren**, das der Hersteller selbst durchführt (Selbstzertifizierung) oder von einer unabhängigen notifizierten Stelle durchführen lässt. Diese Bewertung muss durchlaufen werden, bevor das Produkt – also das AIS – in Verkehr gebracht wird, also vor dem Moment, ab dem sich das Risiko eines AIS manifestieren kann.

Dass der Hersteller die Konformität des Produkts geprüft hat, das anwendbare Konformitätsbewertungsverfahren durchlaufen wurde und die Vorgaben eingehalten sind, wird durch das CE-Kennzeichen angezeigt. Dazu finden sich weitere Informationen im **Blue Guide** der EU-Kommission, dem Leitfaden für die Umsetzung der Produktvorschriften der EU 2022 vom 29. Juni 2022 (https://dtn.re/hrqXlb).

Der AI Act greift diesen Ansatz auf, aber mit einigen Besonderheiten:

- Der AIA reguliert nicht eine Technologie, sondern ihren Einsatz. Er verlangt aber die Einhaltung **grundlegender Anforderungen** an alle HRAIS, gemäss Art. 8-15. Spezifische Use Cases werden durch punktuelle Verbote (→ 27) und durch die Kriterien für die Einstufung als HRAIS (→ 28) festgelegt.

- Die Zuweisung von Pflichten erfolgt über die unterschiedlichen Rollen der Akteure entlang der Wertschöpfungskette (→ 20 ff.). Was im New Approach grundsätzlich die "Hersteller" sind, sind beim AIA die Anbieter, und die "Nutzer" sind die Betreiber.

- Grundsätzlich muss der Anbieter (→ 20) des HRAIS ein **Konformitätsverfahren** durchlaufen, sofern nicht aus besonderen öffentlichen Interessen eine Ausnahme greift (Art. 16 lit. f und Art. 46). Das Konformitätsbewertungsverfahren gibt Art. 43 vor. Der Provider kann für HRAIS im Bereich Biometrie (Anhang III Ziff. 1) wählen, ob er eine Selbstzertifizierung vornimmt (**internes Verfahren**, das Anhang VI festlegt) oder eine notifizierte Stelle (Art. 29 ff. → 56) beizieht (**externes Verfahren**, das Anhang VII festlegt).

  Die Zulässigkeit der Selbstzertifizierung setzt voraus, dass für alle Aspekte des HRAIS harmonisierte Normen (Art. 40) oder gemeinsame Spezifikationen (Art. 41) vorliegen, also harmonisierte Konkretisierungen der grundlegenden Anforderungen bzw. ihrer Umsetzung. Fehlen diese, bleibt dem Provider nur der Weg über eine notifizierte Stelle (Art. 43). Bei den anderen Hochrisiko-Use Cases nach Anhang III gilt generell das Verfahren der Selbstzertifizierung (Art. 43 Abs. 2), und bei HRAIS, die unter eine Produktregulierung nach Anhang I Abschnitt A fallen (bspw. Medizinprodukte), gilt das jeweils anwendbare Verfahren auch für die Konformitätsbewertung nach dem AIA (Art. 43 Abs. 3).

- Der Provider muss für jedes HRAIS eine **EU-Konformitätserklärung** ausstellen und 10 Jahre ab dem Inverkehrbringen oder der In-

betriebnahme des HRAIS zuhanden der Behörden aufbewahren (Art. 16 lit. g und Art. 47). Mit der Konformitätserklärung bringt er zum Ausdruck, dass das HRAIS den entsprechenden Anforderungen entspricht und er dafür verantwortlich ist (Art. 47 Abs. 2 und 4). Die Konformitätserklärung muss die Informationen nach Anhang V enthalten und in eine Sprache übersetzt werden, die für die zuständigen nationalen Behörden "leicht verständlich" ist (Art. 47 Abs. 2).

– Der Provider muss das **CE-Zeichen** anbringen (Art. 16 lit. h und Art. 48). Damit gibt er an, dass er die Verantwortung für die Konformität mit den Anforderungen des AIA und ggf. weiteren anwendbaren Produktanforderungen übernimmt (Art. 30 der Marktüberwachungsverordnung, https://dtn.re/h4EI0Y).

– Das Inverkehrbringen und die Inbetriebnahme sind nicht erlaubt, bevor die Konformitätsbewertung nicht durchlaufen wurde, und bei einer wesentlichen Änderung des HRAIS ist eine erneute Konformitätsbewertung erforderlich (Art. 43 Abs. 5).

– Soweit ein Anbieter **sektorspezifischer** Produktregulierung unterliegt, sind die Anforderungen des AIA generell im entsprechend vorgegebenen Rahmen abzudecken.

– Zudem müssen HRAIS in einer öffentlichen **Datenbank** registriert werden (Art. 49).

HRAIS sind dabei nicht etwa verboten – der AIA ist insofern durchaus innovationsfreundlich. Verboten sind nur wenige Einsatzbereiche oder Use Cases, die als gesellschaftlich besonders unerwünscht taxiert wurden (→ 27).

Umgekehrt sind aber jeweils parallel geltende Anforderungen, Auflagen und Einschränkungen zu beachten, bspw. datenschutz-, lauterkeits-, arbeits- oder immaterialgüterrechtlicher Natur. Der AIA erhält diesbezüglich kaum Erlaubnistatbestände, mit einer Ausnahme beim Datenschutz (→ 1).

16 Wie werden Risiken im AIA kategorisiert?

Der AIA unterscheidet bei der Regulierung unterschiedliche Stufen oder Klassen von Risiken. Massgebend ist dabei vor allem der konkrete Einsatz eines AIS und nicht seine technischen Eigenschaften als solche oder die für das Training oder beim Einsatz verwendeten Daten oder andere Kriterien, die sich für eine Risikoeinstufung ebenfalls anbieten könnten. Diese Differenzierung ist grundsätzlich sinnvoll; allerdings ist sie recht grob und kann den konkreten Umständen nicht in allem gerecht werden, analog zur gesetzlichen Einstufung bestimmter Personendaten als besonders schützenswert. Der AIA kennt vier Risikostufen für AIS: inakzeptables Risiko, hohes Risiko, begrenztes Risiko bzw. Transparenzrisiko und alles andere:

– **Verbotene AIS**: AIS bzw. Use Cases mit inakzeptablen Risiken sind als "verbotene Praxis" generell untersagt (Art. 5 → 27).

– **HRAIS**: AIS bzw. Use Cases in heiklen Bereichen wie kritischen Infrastrukturen, Bildung, Beschäftigung, wesentliche öffentliche Diensten oder Strafverfolgung; sie unterliegen den Anforderungen, die den Hauptteil des AIA ausmachen. Art. 6 regelt die Einstufung eines AIS als HRAIS (→ 28).

– **AIS mit Transparenzrisiken**: Das sind AIS, die zwar keine HRAIS sind, die aber für die direkte Interaktion mit natürlichen Personen bestimmt sind, die Content generieren oder die zur Emotionserkennung oder zur biometrischen Kategorisierung bestimmt sind (Art. 50 → 37). Hier gelten eingeschränkte Anforderungen, die vor allem auf Transparenz zielen.

– **Sonstige AIS**: Für alle anderen AIS enthält der AIS nur am Rande Vorgaben (→ 38).

Die Pflichten einer Risikoklassen gelten dabei jeweils auch für die höheren Klassen.

Prima vista definiert der AIA eine fünfte Risikokategorie: AIS, die "ein Risiko bergen", nach Art. 79. Das sind AIA mit besonderen Risiken nach Art. 3 Nr. 19 der Marktüberwachungsverordnung (https://dtn.re/JgakBQ) also untypisch erhöhten Risiken für die Gesundheit oder Sicherheit oder Grundrechte. Es muss sich dabei nicht um ein HRAIS handeln, auch wenn das i.d.R. der Fall

sein müsste. Hat eine Marktüberwachungsbehörde (→ 43) Grund zur Annahme, dass solche Risiken vorliegen, prüft sie das betreffende AIS und – sollte sich die Annahme bestätigen – informiert die zuständigen nationalen Behörden. Auch Betreiber haben bei einem solchen System besondere Pflichten, nun aber nur dann, wenn es sich um ein HRAIS handelt.

Materiell erhöhen sich die Anforderungen an solche AIS allerdings nicht, es geht nur um eine besondere Prüfung und erforderlichenfalls die Durchsetzung der Compliance. Eine eigene Risikokategorie bilden solche AIS deshalb nicht, und wenn sie nicht zugleich ein HRAIS sind, was allerdings meist der Fall sein dürfte, bestehen kaum Anforderungen.

GPAIM fallen nicht in diese Risikoklassen, weil sie keinen bestimmten Anwendungsbereich haben, der entsprechend klassifiziert werden könnten. Erst wenn sie zu einem GPAIS werden, fallen sie als AIS in eine Risikoklasse.

17 Welche Rollen werden im AIA definiert?

Der AIA definiert mehrere Rollen, die unterschiedliche Pflichten und Verantwortlichkeiten in Bezug auf AIS – und teilweise auch auf GPAI – mit sich bringen. Er folgt dabei mit der Unterscheidung zwischen Anbieter, Betreiber, Einführer und Händler dem Standard des Europäischen Produktesicherheitsrechts, kennt aber auch Rollen:

- **Anbieter** (Provider/AIS und GPAI): Die Stelle (d.h. diejenige natürliche oder juristische Person), die ein AIS in Verkehr bringt und die Hauptverantwortung für die Einhaltung der Anforderungen trägt (→ 20);
- **Betreiber** (Deployer/AIS): Die Stelle, die ein AIS oder eine GPAI einsetzt (→ 21);
- **Einführer** (Importer/AIS): Die Stelle, die ein AIS oder eine GPAI eines Drittstaatproviders erstmals in die EU einführt (→ 23);
- **Händler** (Distributor/AIS): Die Stelle, die ein AIS auf dem Gemeinschaftsmarkt anbietet, ohne selbst ein Anbieter oder Einführer zu sein (→ 24);
- **Produkthersteller** (Product Manufacturer/AIS): Die Stelle, die ein Produkt herstellt, in das ein AIS verbaut wird;
- **Bevollmächtigter** (Representative): Das ist nach Art. 3 Nr. 5 eine Stelle in der EU, die vom Provider schriftlich bevollmächtigt wurde, in seinem Namen die in dieser Verordnung festgelegten Pflichten zu erfüllen bzw. Verfahren durchzuführen. Vertreter haben die Kontroll- und Mitwirkungspflichten nach Art. 22.
- **Betroffene Person**: Die betroffene Person wird nicht legaldefiniert, aber es geht um Personen, deren Daten von einem AIS verarbeitet werden. Sie haben bestimmte Rechte nach dem AIA (zusätzlich zu den Rechten nach der DSGVO).

Hat eine Stelle mehrere Rollen zugleich, gelten die Anforderungen jeweils kumulativ (ErwG 83). ErwG 83 nennt als Beispiel den Händler, der auch Einführer ist, was durch die Legaldefinitionen aber ausgeschlossen ist (ein Händler stellt ein AIS bereit, “mit Ausnahme des Anbieters oder des Einführers”; Art. 3 Nr. 7). Naheliegender ist der Anbieter, der sein AIS in Betreib nimmt und dann auch Betreiber ist.

Zudem definiert der AIA den “**Akteur**” (Operator); das ist ein Oberbegriff für Anbieter, Produkthersteller, Betreiber, Bevollmächtigte, Einführer und Händler (Art. 3 Nr. 8). Er wird im AIA nicht oft verwendet, in der Regel nur zur einfacheren Verweisung und ohne Rechtsfolgen für Akteure zu definieren.

18 Was ist der räumliche Anwendungsbereich des AIA?

Der AIA ist zunächst in der EU anwendbar. Er wird aber ins **EWR**-Recht übernommen und dann auch für Norwegen, Island und Liechtenstein gelten. Derzeit befindet sich der AIA im EWR im Stadium der Prüfung (https://dtn.re/LxZNyE); formell ins EWR-Recht übernommen wird er erst nach einem Beschluss des Draft Joint Committee.

Wie die DSGVO will der AI Act einen gewissen Grundschutz und ein Level Playing Field innerhalb des EWR festlegen (ErwG 22). Er muss deshalb auch bestimmte Fälle mit inter-regionaler Komponente erfassen. Dabei unterscheidet der AIA zwischen den einzelnen Rollen in der Wertschöpfungskette, weshalb → 17 vorangestellt wurde.

Nach Art. 2 und 3 (beide Bestimmungen sind zusammen für den Anwendungsbereich massgebend) findet er in räumlich-persönlicher Hinsicht wie folgt Anwendung:

- Für **Anbieter** (Provider):
  - unabhängig vom Standort des Providers dann, wenn ein AIS oder ein GPAIM in der EU in Verkehr gesetzt oder in Betrieb genommen wird (Art. 2 Abs. 1 lit. a); und
  - wenn der Output des Systems in der EU verwendet wird (lit. c → 19);
- für **Betreiber** (Deployer):
  - wenn der Deployer in der EU niedergelassen sind bzw. sich in der EU befinden (lit. b). Die "Niederlassung" dürfte analog zur DSGVO weit ausgelegt werden;
  - wenn der Output des Systems in der EU verwendet wird (wiederum lit. c);
- für **Einführer** (Importer): wenn er in der EU ansässig ist und ein AIS einführt (Art. 3 Nr. 6);
- für **Händler** (Distributor): wenn das AIS auf den auf dem EU-Markt bereitgestellt wird, unabhängig vom Standort des Händlers (Art. 3 Nr. 7);
- für **Produkthersteller** (Manufacturer): wenn sie ein AIS zusammen mit ihrem Produkt in eigenen Namen in der EU in Verkehr bringen oder in Betrieb nehmen (Art. 2 Abs. 1 lit. e);
- für (EU-)**Vertreter** ausländischer Provider (Art. 2 Abs. 1 lit. f);
- für **betroffene Personen** in der EU (Art. 2 Abs. 1 lit. g).

Ein schweizerisches Unternehmen kann also insbesondere dann unter den AIA fallen, wenn es:

- ein AIS in der bzw. in die EU verkauft (als Entwickler, Einführer oder Händler),
- ein anderes Produkt in der EU verkauft, das ein AIS als Komponente verwendet,
- Output generiert, der in der EU verwendet wird (→ 19).

19 Was bedeutet "Output wird in der EU verwendet"?

Output wird in Art. 2 Abs. 1 lit. c umschrieben:

> *c) Anbieter und Betreiber von KI-Systemen, die ihren Sitz in einem Drittland haben oder sich in einem Drittland befinden, wenn die vom KI-System hervorgebrachte Ausgabe in der Union verwendet wird;*

Dazu gehört sicher etwa AI-generierter Text oder ein Bild. Allerdings enthält der AIA keine eigene Definition der hervorgebrachten Ausgabe, im Gegensatz zum Input (Art. 3 Nr. 33). Der Begriff wird häufiger verwendet, aber jeweils ohne nähere Umschreibung (bspw. in ErwG 12 bei der Definition des AIS → 13).

An einigen Stellen wird Output aber in einer Weise verwendet, die eine **breite Auslegung** nahelegt, sofern man eine einheitliche Verwendung dieses Begriffs unterstellt (etwa in Anhang III Ziff. 8 lit. b, HRAIS beim Einsatz für die Beeinflussung einer Wahl oder Abstimmung: ein AIS ist hier nicht als HRAIS erfasst, wenn sein Output nicht direkt natürliche Personen betrifft wie etwa bei einem Tool zur Kampagnenorganisation: Hier kann Output nicht nur das Ergebnis von Generative AI meinen). Es liegt deshalb und aufgrund des Schutzzwecks des AIA nahe, auch etwa AI-generierte Steuerungssignale unter den Begriff des Output zu fassen.

Wichtiger ist daher die Frage, wann Output in der EU verwendet wird. Jeder Spillover kann nicht gemeint sein. Man wird vielmehr eine gewisse **Spürbarkeit der Auswirkung** in der EU verlangen müssen, die analog zu Marktverhaltensregeln nur durch eine Ausrichtung konkretisiert werden kann. Dafür spricht insbesondere auch ErwG 22, der eine Umgehung verhindern, aber nicht jeden beliebigen Effekt in der EU erfassen will, von der "Absicht" spricht und als Beispiel eine Konstellation nennt, bei der klar nicht nur ein Spillover vorliegt:

> *Um die Umgehung dieser Verordnung zu verhindern [...], sollte diese Verordnung auch für Anbieter und Betreiber von KI-Systemen gelten, die in einem Drittland niedergelassen sind, soweit beabsichtigt wird, die von diesem System erzeugte Ausgabe in der Union zu verwenden.*

und:

> *Dies ist beispielsweise der Fall, wenn ein in der Union niedergelassener Akteur bestimmte Dienstleistungen an einen in einem Drittland niedergelassenen Akteur im Zusammenhang mit einer Tätigkeit vergibt, die von einem KI-System ausgeübt werden soll [...]. Unter diesen Umständen könnte das von dem Akteur in einem Drittland betriebene KI-System [...] dem vertraglichen Akteur in der Union die aus dieser Verarbeitung resultierende Ausgabe dieses KI-Systems liefern [...]*

Ein Anbieter kann deshalb alleine aufgrund einer Verwendung des Outputs solange nicht unter den AIA fallen, als der Output nicht auf eine Verwendung in der EU ausgerichtet ist, d.h. bestimmungsgemäss in der EU verwendet wird. Eine gewisse Konkretisierung kann dabei über den Blue Guide (→ 15) erfolgen, der allerdings vage bleibt.

Der Anwendungsbereich ist auch so breit genug. Wenn ein Mitarbeiter eines schweizerischen Unternehmens eine E-Mail an eine französische Kollegin schickt und darin AI-generierten Text verwendet, oder wenn eine Präsentation mit einem AI-generierten Bild oder ein durch AI transkribiertes Protokoll an einen Empfänger in der EU versandt wird, dürfte dies ausreichen, es sei denn, man leite aus dem Kriterium der Spürbarkeit eine Bagatellschwelle ab, die zusätzlich zum Erfordernis der Ausrichtung zur Anwendung käme.

Die Frage muss vorläufig allerdings offenbleiben – es ist anzunehmen, dass das EAIB (→ 53) hier Konkretisierungen vorschlagen wird. Für Nicht-EU-Akteure, die nur Betreiber eines HRAIS sind, und für Akteure, die sich nur mit Nicht-Hochrisiko-AIS befassen, spielt diese Frage allerdings eine nicht ganz so wesentliche Rolle wie für HRAIS-Anbieter.

Aufgrund der Legaldefinition des Anbieters kann man zudem die Frage stellen, ob die Verwendung des Output alleine überhaupt ausreichen kann oder ob **zusätzlich auch ein Inverkehrbringen oder eine Inbetriebnahme** in der EU vorausgesetzt wird. Gegen diese Auslegung sprechen aber mehrere Argumente:

- ErwG 22 weitet den Anwendungsbereich für KIS aus, "selbst wenn sie in der Union weder in Verkehr gebracht noch in Betrieb genommen oder verwendet werden".
- Mit Bezug auf Anbieter müsste Art. 2 die Verwendung von Output bei dieser Auslegung gar nicht mehr erwähnen, weil das Inverkehrbringen in der EU alleine schon genügt (Art. 2 Abs. 1). Beim Betreiber dagegen hätte der Hinweis auf den Output auch dann eine Berechtigung, wenn beim Provider das Inverkehrbringen bzw. die Inbetriebnahme verlangt werden.
- Die enge Auslegung würde zur Situation führen, in der ein Betreiber dem AIA unterliegen kann, nicht aber der Anbieter des entsprechenden Systems. Da die Pflichten des Betreibers zumindest in Teilen voraussetzen, dass auch der Anbieter seinen Pflichten nachgekommen ist (bspw. bei der Aufbewahrung

von Log-Daten, die nicht möglich ist, wenn der Anbieter nicht für die Logfähigkeit des HRAIS gesorgt hat), liegt ein Gleichlauf näher.

– Die Legaldefinition des Anbieters lässt den Schluss zu, dass ein Inverkehrbringen bzw. eine Inbetriebnahme nur dann eine Voraussetzung für die Anbietereigenschaft ist, wenn eine Stelle ein AIS nicht selbst entwickelt, sondern entwickeln lässt. Bei selbstentwickelter AIS genügt nach dieser Auslegung schon die Entwicklung des AIS (→ 20).

– Aus Schutzüberlegungen werden Behörden und Gerichte wohl einer weiten Auslegung folgen, den Output also genügen lassen. Dafür sprechen jedenfalls die Erfahrungen mit der grundrechtsbezogenen Auslegung der DSGVO.

Bis zur Klärung der Frage sollte daher davon ausgegangen werden, dass die bestimmungsgemässe Verwendung des Output ausreicht genügt.

Fragen kann man sich allerdings, ob auch eine **Verwendung *als* Output** erforderlich ist. Das dürfte zutreffen: Wer AI-generierte Texte in der EU verwendet wissen will, wird zwar auch dann unter den AIA fallen können, wenn er einen Screenshot mit dem entsprechenden Text verwendet. Wer dagegen zu Texte generiert, um die Funktionsweise eines LLM zu illustrieren und generierte Texte als Beispiele verwendet und nicht aufgrund ihres eigentlichen Aussagegehalts, verwendet kaum Output in der EU.

Rollen

20 Was ist ein Anbieter (Provider)?

Die "Anbieter" haben die Rolle, die im Produktsicherheitsrecht den "Herstellern" zukommt. Es sind diejenigen Stellen, die AIS oder ein GPAIM entwickeln (oder für sich unter ihrer Kontrolle entwickeln lassen) und in Verkehr bringen oder in Betrieb nehmen (Art. 3 Nr. 3):

> *[...] eine [...] Stelle, die ein KI-System oder ein KI-Modell mit 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;*

Anbieter tragen die **Hauptverantwortung** für die Konformität des AIS, bspw. durch das Konformitätsbewertungsverfahren, das Risikomanagement, Gewährleistung der Datenqualität beim Training und die Überwachung nach dem Inverkehrbringen (→ 0).

Die Formulierung von Art. 3 Nr. 3 lässt allerdings **zwei Auslegungen** zu:

- Die Voraussetzung, dass ein AIS in Verkehr gebracht oder in Betrieb genommen wird, kann generell gelten
- oder aber nur für den zweiten Fall, bei dem ein AIS nicht selbst entwickelt wird ("entwickeln lässt").

Auf den ersten Blick ist die erste Auslegung naheliegender. Eindeutig ist das aber keineswegs. Für die räumliche Anwendung genügt eine Verwendung des Output in der EU (→ 19). Es wäre widersprüchlich, die meisten Pflichten entfallen zu lassen, weil die betreffende Stelle das verwendete (HR)AIS nicht auch in der EU in Verkehr bringt oder in Betrieb nimmt. Mit anderen Worten: Diese weite Auslegung des Anbieterbegriffs löst den inneren Widerspruch bei Art. 2 auf, denn dann genügt die Verwendung des Output in der EU eindeutig. Dies spricht für die weitere Auslegung des Anbieterbegriffs, wie er auch in der Literatur vertreten wird.

"**Inverkehrbringen**" ("Placing on the market"; AIS oder GPAIM) wird in Art. 3 Nr. 9 definiert als der Vorgang, durch den ein bestimmtes AIS oder ein bestimmtes GPAIM erstmals im Unionsmarkt bereitgestellt wird:

- Das kann einmalig oder dauerhaft geschehen, für jedes einzelne AIS oder GPAIM aber nur einmal. Wer ein AIS einem Kunden in der EU zur Verfügung stellt, wird also nicht zum Anbieter, wenn das AIS bereits in der EU in Verkehr gebracht ist.
- Das Inverkehrbringen meint dabei ein Angebot oder eine Vereinbarung zur Übertragung des Eigentums, des Besitzes oder sonstiger Rechte am AIS bzw. GPAIM voraus, entgeltlich oder unentgeltlich. Bei einem AIS ist das bspw. dann der Fall, wenn ein AIS zur Verwendung *on premise* oder als SaaS-Angebot überlassen wird, bspw. über eine Schnittstelle (API; vgl. ErwG 97 und Art. 6 der Marktüberwachungsverordnung zum Fernabsatz). Das Inverkehrbringen erfolgt durch den Anbieter oder – im Fall eines AIS – einen Einführer (Importeur; s. unten). Geben diese ein AIS an einen Vertreiber (Distributor) für den weiteren Vertrieb weiter, bringen sie das AIS bereits in Verkehr (die folgende Handlung des Vertreibers ist dann eine "Bereitstellung").
- Kein Inverkehrbringen wäre dagegen die Einfuhr durch eine Person für den Eigengebrauch, also bspw. eines Handys mit AI-Anwendungen, die Übergabe eines AIS zu reinen Testzwecken oder die Demonstration eines AIS an einer Fachmesse (vgl. den Blue Guide, Ziff. 2.3).

Das "**in Betrieb nehmen**" ("Putting into service"; AIS) wird in Art. 3 Nr. 11 sodann definiert als der

Vorgang, bei dem ein AIS dem Betreiber (Deployer) für dessen erstmalige Verwendung abgegeben wird, aber auch die eigene Erstverwendung durch den Anbieter (Provider):

- Wer ein AIS entwickelt und selbst einsetzt, ist ein Anbieter im Sinne des AIA mit den entsprechenden Pflichten.
- Auch Betreiber, Einführer, Vertreiber oder sonstige Stellen können nachträglich zum Anbieter werden (→ 22).

Weil ein Produkt durch den Einbau eines AIS ("**Embedded AIS**") nicht selbst zum AIS wird, wird der Hersteller des entsprechenden Produkts auch nicht zum Anbieter i.S.d. AIA, sofern das Embedded AIS unter dem Namen bzw. der Marke einer anderen Stelle zur Anwendung kommt.

Bei einer **Kombination von AIS** dürfte ebenfalls jeder einzelne Anbieter als Anbieter gelten, sofern die Komponenten weiterhin bestimmungsgemäss verwendet werden. Weil der AIA aber auf "Systeme" und nicht Softwarepakete Bezug nimmt, können Komponenten wohl gemeinsam als AIS gelten, wenn sie funktional eine Einheit bilden.

Hersteller eines regulierten Produkts, das unter eine Produktregulierung nach Anhang I fällt, weil ein AIS als Sicherheitsbauteil (i.S.v. Art. 3 Nr. 14) verbaut wurde, und die das Produkt im eigenen Namen in Verkehr bringen oder in Betrieb nehmen, gelten sodann ebenfalls als Anbieter (Art. 25 Abs. 3).

21 Was ist ein Betreiber (Deployer)?

Betreiber gestalten das System nicht selbst, sie setzen es bloss ein (Art. 3 Nr. 4) – nach dem allgemeinen Produktsicherheitsrecht sind es also "Endnutzer".

Allerdings muss die Verwendung des AIS "**under the authority**" des Betreibers erfolgen, "in eigener Verantwortung" (Art. 3 Nr. 4). Das setzt voraus, dass das System nicht alleine im Auftrag eines anderen Betreibers betrieben wird. Offen ist, ob es auch verlangt, dass der Betreiber das AIS selbst konfiguriert, steuert, parametrisiert usw. oder ob es genügt, dass er selbst über die Verwendung entscheidet. Geht man von den Pflichten des Betreibers aus und stellt man die Frage, wann diese Pflichten greifen können, genügt schon eine niedrigere Schwelle, der blosse Einsatz ohne weitergehende Kontrolle wäre hier nicht vorausgesetzt. "Under it's authority" heisst nach dieser naheliegenden Sicht, dass der Einsatz nicht alleine im Sinne einer Auftragsbearbeitung oder durch einen Arbeitnehmer erfolgt, sondern durch eine Stelle, die ein AIS für ihre eigenen Zwecke einsetzt. Wer ein AIS für einen anderen einsetzt, ist umgekehrt also kein Betreiber (aber meist ein Anbieter).

Der Betreiber muss sich an die **Betriebsanleitung** halten (→ 35). Diese ist deshalb wesentlich, weil diese u.a. den bestimmungsgemässen Gebrauch des AIS bestimmt, d.h. die "Zweckbestimmung" (Art. 3 Nr. 12), für die das AIS bestimmt ist, ebenso wie den Rahmen des korrekten Einsatzes. Verlässt der Betreiber diesen Rahmen, kann er zum Anbieter werden (→ 22).

Bei GPAIM fehlt ein Betreiber, weil ein GPAIM nicht betrieben werden kann (→ 39).

22 Wann wird der Betreiber zum Anbieter?

Diese Frage ist weniger einfach zu beantworten, als es zunächst scheint. Art. 25 AIA enthält die Grundregel, dass ein Betreiber unter bestimmten Umständen ein Anbieter wird (sog. "deemed provider"):

- wenn er **als Anbieter auftritt**, indem er seinen Namen oder seine Marke auf einem HRAIS anbringt, nachdem dieses vom ursprünglichen Anbieter in Verkehr gebracht oder in Betrieb genommen wurde,

- wenn er das HRAIS **wesentlich ändert** (wie in Art. 3 Nr. 23 AIA definiert), aber ohne das HRAIS dadurch zum low-risk-AIS zu machen, und
- wenn er ein AIS ausserhalb seiner Zweckbestimmung so einsetzt, dass er es erst **zum HRAIS macht**.

Dabei gilt jeweils nur der "deemed provider" als Anbieter; der ursprüngliche Anbieter wird insoweit aus seiner Verantwortung entlassen. Er muss aber mit dem neuen Anbieter zusammenarbeiten (Art. 25 Abs. 2). Das kann er wohl entsprechend bepreisen. Die Kooperationspflicht entfällt allerdings, wenn der Erstanbieter vorgegeben hat, dass das AIS nicht in ein HRAIS umgewandelt werden darf – auch dies spricht daher für eine entsprechende **Vertragsgestaltung**.

Für eine Einstufung als Anbieter nicht genügend ist demgegenüber der blosse Einsatz eines HRAIS **ausserhalb der bestimmungsgemässen Verwendung**. Der Anbieter muss mit einer solchen vielmehr bis zu einem gewissen Grad rechnen, wie neben Art. 25 auch Art. 9 Abs. 2 lit. b zeigt: Das RMS des Anbieters muss auch den Risiken bei vorhersehbaren Missbrauchsfällen Rechnung tragen. Erst wenn der Missbrauch zu einer wesentlichen Änderung führt oder ein AIS erst zum HRAIS macht, wird der Betreiber nach Art. 25 zum "deemed provider". Wer einen für den Kundensupport bestimmten Chatbot zur Auswahl von Stellenbewerbern einsetzt, wird deshalb zum HRAIS-Anbieter, nicht aber beim Einsatz für Mitarbeiter-Zufriedenheitsumfragen (kein HRAIS).

Auch ein **Finetuning** (→ 12) sollte nicht ausreichen, um zum Anbieter des entsprechend weiter trainierten AIS zu werden, es sei denn, der Betreiber biete das AIS unter eigenem Namen an oder setze es in einer Weise ein, dass es neu zum HRAIS wird. Es ist zwar offen, ob sich die Qualifikation des Anbieters im Fall eines Finetuning an Art. 25 anlehnt oder schlicht auf das nicht näher definierte Element des "Entwickelns" nach Art. 3 Nr. 3 abstellt. Im letzteren Fall könnte der Betreiber im Fall eines Finetunings eher als Anbieter eingestuft werden. Allerdings verwendet der AIA den Ausdruck "entwickeln" ("develop") in der Regel in einem weitergehenden Sinn (bspw. in Art. 2 Abs. 6: keine Anwendung des AIA auf ein AIS, das alleine für Forschungszwecke "entwickelt" [und in Betrieb genommen] wurde). Zudem grenzt ErwG 93 den Bereich der Entwicklung von der Rolle des Betreibers ab. Vor allem aber dürfte ins Gewicht fallen, dass der Anwender im Falle eines Finetunings die Pflichten des Anbieters kaum erfüllen kann, weil seine Kontrolle des AIS nicht weit genug geht. Kein Anbieter wird der Betreiber eines GPAIS nur dadurch, dass er ein RAG (→ 12) verwendet.

Bei **GPAIM** gilt weiter, dass das Modell zum GPAIS wird, sobald das Modell als Produkt bereitgestellt wird, und sei es nur dadurch, dass es mit einer Nutzerschnittstelle ergänzt wird (→ 39). Anschliessend gelten die vorstehenden Vorgaben. Wer also ein GPAIM einkauft und dann für einen bestimmten Use Case in Betrieb nimmt, ist Anbieter des resultierenden AIS.

23 Was ist ein Einführer (Importer)?

Der Einführer ist eine Stelle in der EU, die ein fremdes HRAIS (d.h. ein unter fremdem Namen bzw. fremder Marke angebotenes HRAIS) in die EU einführt (Art. 3 Nr. 6).

Der Einführer muss die Konformität nicht selbst herstellen, aber seine Pflichten bauen auf jenen des Anbieters auf – er ist mit anderen Worten nicht bloss Reseller, sondern muss

- kontrollieren, dass die Konformitätsbewertung erfolgt ist, die technische Dokumentation gemäss Art. 11 und Anhang IV AIA vorliegt, das HRAIS das CE-Kennzeichen trägt und der Anbieter einen Bevollmächtigten bestellt hat (Art. 23 Abs. 1), und
- die Dokumentation zuhanden der Aufsichtsbehörden aufbewahren (Abs. 5).
- Bestehen Zweifel an der Einhaltung der grundlegenden Anforderungen, darf das HRAIS nicht in Verkehr gebracht werden, und

– bei höheren Risiken (i.S.v. Art. 79 Abs. 1) müssen der Anbieter, der Bevollmächtigte und die zuständigen Marktüberwachungsbehörden entsprechend informiert werden (Art. 23 Abs. 2).

24 Was ist ein Händler (Distributor)?

Das ist nach Art. 3 Nr. 7 eine Stelle, die ein HRAIS von Anbieter, einem Einführer oder einem anderen Händler bezieht und auf dem Unionsmarkt bereitstellt, ohne selbst ein Anbieter oder Einführer zu sein, d.h. nach dem Inverkehrbringen. “Bereitstellen” meint dabei jede entgeltliche oder unentgeltliche Abgabe eines AIS oder einer GPAI zum weiteren Vertrieb oder zur Verwendung auf dem Unionsmarkt (Art. 3 Nr. 10).

Ähnlich wie der Einführer muss der Händler

– prüfen, dass das HRAIS das CE-Kennzeichen trägt, dass eine Konformitätserklärung und die Betriebsanleitung vorliegen und dass der Anbieter oder auch der Einführer ihren Namen bzw. ihre Marke angegeben haben und über ein QMS (→ 35) verfügen.
– Bei berechtigten Zweifeln an der Einhaltung der grundlegenden Anforderungen darf das HRAIS wiederum nicht bereitgestellt werden, und der Händler muss mit dem Anbieter oder Einführer Kontakt aufnehmen.
– Können Mängel nicht behoben werden, muss das HRAIS vom Markt genommen oder zurückgerufen werden (vom Händler, von Anbieter oder vom Einführer; Art. 24 Abs. 4).
– Bei höheren Risiken (nach Art. 79 Abs. 1) müssen der Anbieter oder der Einführer und die zuständigen Behörden informiert werden (Art. 24 Abs. 4).

25 Was ist ein Produkthersteller (Product Manufacturer)?

Auch diese Rolle wird nicht legaldefiniert. Es handelt sich um eine Stelle, die ein Produkt herstellt, in das ein AIS integriert wird. Unter bestimmten Umständen wird diese Stelle dadurch zum Anbieter, nämlich dann, wenn das AIS eine **Sicherheitskomponente** ihres Systems ist, dieses unter eine Produktregulierung nach Anhang I fällt und der Produkthersteller das AIS als Teil des eigenen Produkts im eigenen Namen auf dem Markt bereitstellt bzw. das Produkt nach der Bereitstellung auf dem Markt im Namen des Produktherstellers in Betrieb genommen wird (Art. 25 Abs. 3). In diesem Fall muss der Produkthersteller sicherstellen, dass das verbaute AIS den Anforderungen entspricht (ErwG 87).

26 Wann muss ein Bevollmächtigter in der EU bestellt werden?

Nach Art. 22 muss der **Anbieter** eines HRAIS einen Bevollmächtigten bestellen, wenn er ausserhalb der EU niedergelassen ist. Ein “Bevollmächtigter” ist nach Art. 3 Nr. 5 eine in der EU ansässige oder niedergelassene Stelle, die der Anbieter eines AIS oder eines GPAI-Modells schriftlich (d.h. wohl in Textform) bevollmächtigt hat und sich einverstanden erklärt, in dessen Namen die Pflichten nach dem AIA zu erfüllen bzw. Verfahren durchzuführen.

Die Aufgaben des Bevollmächtigen sind vertraglich festzulegen, umfassen aber mindestens den Katalog nach Art. 22 Abs. 3, bspw. die Prüfung, ob die Konformitätserklärung und die technische Dokumentation erstellt wurden und das Konformitätsbewertungsverfahren durchgeführt wurde, die Bereithaltung bestimmter Angaben und Unterlagen zuhanden der Behörden und Mitwirkungspflichten bei der Registrierung des HRAIS. Eine analoge Bestimmung enthält Art. 54 für Anbieter eines GPAIM (→ 39).

Bevollmächtige können ihr Mandat niederlegen, und u.U. müssen sie das sogar.

Betreiber und andere Akteure ausser dem Anbieter haben keine Pflicht, einen Bevollmächtigen zu bestellen.

Verbotene und hochriskante Anwendungen

27 Welche Anwendungsfälle sind verboten?

AIS bzw. Use Cases mit inakzeptablen Risiken sind als "verbotene Praxis" ausnahmsweise untersagt, d.h. untersagt ist jeweils das Inverkehrbringen, die Inbetriebnahme oder die Verwendung eines AIS für einen entsprechenden Zweck (Art. 5):

- **Unterschwellige Beeinflussung** (Art. 5 Abs. 1 lit. a): Manipulation, die das Verhalten unbewusst beeinflussen, dabei eine Entscheidung verfälschen und so einen Schaden verursachen. Dabei geht es bspw. um Formen der Täuschung bspw. durch "Dark Patterns" oder auch ein "Nudging", besonders durch ein so niederschwelliges Vorgehen, dass es nicht bewusst wahrgenommen wird, bspw. in einer virtuellen Umgebung (ErwG 29). Eine Täuschungsabsicht ist dabei nicht grundsätzlich vorausgesetzt, weil die absichtliche Täuschung nur eine Tatbestandsvariante ist.
- **Ausnutzen von Schutzbedürftigkeit**, aufgrund von Alter, Behinderung, etc. (Art. 5 Abs. 1 lit. b). Auch hier geht es um die schädliche Verfälschung von Entscheidungen (ErwG 29). Nicht erfasst ist eine verhältnismässige Affirmative Action;
- **Social Scoring** (Art. 5 Abs. 1 lit. c): Bewertung von Personen nach sozialem Verhalten oder persönlichen Eigenschaften über längere Zeiträume, wenn Personen dabei unfair behandelt werden, d.h. wenn der Einsatz des AIS für betroffene Personen eine unerwartete oder unverhältnismässige Folge hätte. Nicht erfasst ist dabei die Bonitätsbewertung, die nicht verboten, sondern hochriskant ist (→ 32);
- **Risikobewertung für Straftaten** (Predictive Policing) durch Profiling (Art. 5 Abs. 1 lit. d; mit Ausnahmen);
- **Gesichtserkennung**: Erstellen von Gesichtserkennungsdatenbanken durch breites Scraping von Bildern aus dem Internet oder Überwachungsaufnahmen (Art. 5 Abs. 1 lit. e). Nicht erfasst wäre bspw. der Abgleich eines Bildes mit Bildern im Internet, weil es dabei nicht zu einem Scraping kommt;
- **Emotionserkennung** am Arbeitsplatz oder in Bildungseinrichtungen (Art. 5 Abs. 1 lit f; mit Ausnahmen für gesundheits- oder sicherheitsbezogene Anliegen). Die Emotionserkennung in anderen Bereichen ist nicht verboten. Verboten wäre bspw. eine Transkription von Calls mit einer Auswertung, ob ein Kundenberater ausreichend freundlich ist oder ob ein Mitarbeiter negative Emotionen gegenüber dem Unternehmen äussert. Weil der AIA bei diesem Verbot nicht mit dem definierten Begriff des Emotionserkennungssystems arbeitet, ist eine Erkennung von "Absichten" (Art. 3 Nr. 39) nicht erfasst, es muss um Emotionen gehen, aber Grundlage einer Absichtserkennung können nicht nur biometrische, sondern auch andere Daten sein;
- **Kategorisierung nach biometrischen Daten**, um auf Rasse, politische Einstellungen, religiöse Überzeugungen, sexuelle Orientierung usw. zu schliessen (Art. 5 Abs. 1 lit. g; mit Ausnahmen). Der Begriff der "biometrischen Daten" wird in Art. 3 Nr. 34 definiert, es muss um Personendaten gehen. Ausgenommen vom Verbot sind AIS aber dann, wenn die Kategorisierung nur eine aus objektiven technischen Gründen notwendige Nebenfunktion eines anderen kommerziellen Dienstes ist (Art. 3 Nr. 40); bspw. wenn ein Online-Dienst für den Kleiderkauf Körpermerkmale verwendet (soweit es sich dabei um biometrische Daten handelt);
- **Realtime-biometrische Fernidentifizierung** in öffentlich zugänglichen Bereichen (Art. 5 Abs. 1 lit. h und Abs. 2-7; mit Ausnahmen). Nicht erfasst ist eine Authentifizierung (→ 29).

Die Kommission zudem Leitlinien zu den verbotenen Praktiken erlassen (→ 51).

Diese Verbote können sich mit weiteren Verboten überschneiden, bspw. lauterkeitsrechtlichen Täuschungsverboten oder datenschutzrechtlichen Schranken. Dass ein AIS nicht verboten ist, heisst also nicht im Umkehrschluss, dass es generell erlaubt ist. Schranken können sich bspw. aus dem Datenschutz- und dem Lauterkeitsrecht ergeben.

28 Was ist ein Hochrisiko-AI-System?

AIS bzw. Use Cases in heiklen Bereichen wie kritischen Infrastrukturen, Bildung, Beschäftigung, wesentliche öffentliche Diensten oder Strafverfolgung; sie unterliegen den Anforderungen, die den Hauptteil des AIA ausmachen (→ 15). Art. 6 regelt die Einstufung eines AIS als HRAIS.

Dabei sind zwei Fälle zu unterscheiden:

Der erste Fall nach Art. 6 Abs. 1 betrifft AIS, die unter eine **Produktregulierung nach Anhang I** fällt, weil das AIS bzw. sein Use Case selbst unter eine solche Regulierung fällt oder weil es als Sicherheitsbauteil (i.S.v. Art. 3 Nr. 14) in ein solches Produkt verbaut wurde). Hier steht das Produktrisiko im Vordergrund, insbesondere Risiken für Leib und Leben. Der Anhang I unterscheidet dabei zwei Kategorien:

- Die erste Kategorie in Abschnitt A betrifft Produktregulierungen, die dem **New Approach** folgen. Hier ist der AIA direkt anwendbar. Das betrifft bspw. Maschinen, Spielzeuge, Explosivstoffe oder Medizinprodukte.
- Die zweite in Abschnitt B betrifft Produktregulierungen ausserhalb des New Approach. Der AIA ist hier **nicht direkt anwendbar**. Stattdessen werden die entsprechenden Rechtsakte in Art. 102 ff. so angepasst, dass die Vorgaben aus dem Kapitel III Abschnitt 2 (Art. 8 ff., grundlegende Anforderungen an HRAIS) im Sektorerlass berücksichtigt werden. Das betrifft Transportmittel (Luftfahrt, Eisenbahn, Kraftfahrzeuge usw.).

Vorausgesetzt ist dabei jeweils, dass das Produkt bzw. AIS als Produkt die Durchführung einer **Konformitätsbewertung** durch Dritte verlangt (Art. 6 Abs. 1 lit. b). Ob dies auch Fälle erfassen kann, bei denen ein internes Konformitätsbewertungsverfahren zur Anwendung kommt, ist strittig.

Der zweite Fall nach Art. 6 Abs. 2 betrifft AIS, die im **Anhang III** genannt wird. Anhang III betrifft bestimmte Einsatzgebiete; Anknüpfungspunkt ist hier also weniger ein Produkt- als vielmehr ein Einsatzrisiko. Es geht um die folgenden, abschliessend aufgezählten Fälle, wobei es jeweils um den bestimmungsgemässen Einsatz des HRAIS geht (näher 29 ff.):

- **Biometrie**: Einsatz von AIS zur biometrischen Fernidentifizierung, biometrischen Kategorisierung oder Gefühlserkennung (vgl. → 27);
- **Kritische Infrastruktur**: AIS, die als Sicherheitskomponente in bestimmten kritischen Infrastrukturen dient (→ 31);
- **Allgemeine und berufliche Bildung**: AIS zur Steuerung des Zugangs zu Bildungsangeboten, der Bewertung von Lernergebnissen oder der Überwachung bei Prüfungen (→ 30);
- **Beschäftigung, Personalmanagement und Zugang zur Selbständigkeit**: AIS im Bereich Recruiting oder für relevante Entscheidungen oder die Beobachtung und Bewertung von Leistung oder Verhaltens (→ 30);
- **Grundlegende Dienste und Leistungen**: AIS zur Beurteilung des Anspruchs auf öffentliche Unterstützung (bspw. Sozialversicherung), Bonitätsbeurteilung, Risiko- und Prämienbestimmung in der Lebens- und Krankenversicherung oder Triage von Notrufen, Notfalleinsätzen und der ersten Hilfe (→ 32);
- AIS zur Unterstützung von Strafverfolgungsbehörden, im Bereich Migration, Asyl und Grenzkontrolle und in der Justiz und der demokratischen Meinungsbildung (→ 33).

Massgebend ist dabei die bestimmungsgemässe Verwendung des AIS, wobei die Zweckbestimmung entweder vom Hersteller gesetzt wird (Art. 3 Nr. 12) oder dann vom Betreiber, der ein AIS ausserhalb des Zweck einsetzt (Art. 25 → 22).

29 Welche Fälle sind im Bereich der Biometrie hochriskant?

Anhang III Ziff. 1 regelt Use Cases im Bereich der Biometrie. Drei Fälle sind erfasst:

- Der erste Fall ist die **biometrische Fernidentifizierung**. Diese wird in Art. 3 Nr. 41 legaldefiniert. Es geht um AIS, die dafür bestimmt sind, Personen ohne deren Mitwirkung und in der Regel aus der Ferne zu identifizieren. Nicht erfasst sind damit Authentifizierungssysteme bei Räumlichkeiten und Geräten wie z.B. Iris-, Gesichts-, Venen- und Fingerabdruckscanner (s. auch ErwG 54). Erfasst wäre dagegen eine über einer Autobahn montierte Kamera, wenn ein AIS die Bilder mit einer Datenbank abgleicht.

- Der zweite Fall betrifft die **biometrische Kategorisierung** von Menschen, wenn ein AIS dazu bestimmt ist, auf "sensible oder geschützte Attribute" zu schliessen (bspw. Menschen AI-gestützt in Ethnien eingeteilt werden). Nicht erfasst sind wiederum (→ 27) Fälle, bei denen die Kategorisierung nur eine aus objektiven technischen Gründen notwendige Nebenfunktion eines anderen kommerziellen Dienstes ist (Art. 3 Nr. 40).

- Der dritte Fall sind AIS zur **Emotionserkennung**. Das sind nach Art. 3 Nr. 39 AIS, die "Emotionen oder Absichten" feststellen bzw. prognostizieren sollen, dies aber auf Grundlage biometrischer Daten. Das betrifft bspw. ein AIS, das aus der Stimme – Färbung, Zittern etc. – auf Emotionen schliesst. Auch ein Schluss auf die Gesundheit dürfte erfasst sein, in weiter Auslegung. Die Grundlage muss aber ein biometrisches Datum sein. Werden Emotionen (oder Absichten) auf Basis von E-Mails oder anderen Texten eingeschätzt, wird das AIS dadurch nicht zum HRAIS. Allerdings: Im Arbeitsbereich kann das Ergebnis anders ausfallen, denn hier wird ein AIS u.a. dann zum HRAIS, wenn es dazu dient, Entscheidungen über Arbeitsbedingungen, Beförderung, Kündigung usw. zu beeinflussen, oder Leistung oder Verhalten zu beobachten (→ 30). Das gilt selbstverständlich auch dann, wenn die Inputdaten biometrische Daten sind.

30 Welche Fälle sind im Arbeits- und Bildungsbereich hochriskant?

Anhang III nennt wie erwähnt Anwendungsfälle (Use Cases), die als hochriskant gelten (→ 28). Anhang III Ziff. 3 betrifft die **berufliche und ausserberufliche (Weiter-)Bildung:**

- Ein erster Anwendungsfall (lit. a) sind AIS, die eingesetzt werden sollen, um den **Zugang oder die Zulassung zu Bildungsangeboten** festzustellen. "Feststellen" meint "Bestimmen", wie sich aus ErwG 56 ergibt – hochriskant ist also ein AIS, dessen bestimmungsgemässe Verwendung eine Gatekeeper-Funktion für Bildungsangebote ist, bspw. bei einer Zulassungs- oder Eignungsprüfung. Das betrifft nicht nur Entscheidungen über den Zugang als solchen, sondern auch die Auswahl aus unterschiedlichen Bildungsangeboten. "Bestimmen" ist dabei mehr als "dabei mitwirken". Ein AIS, das Empfehlungen für den Zugang abgibt, wäre daher nicht wenig erfasst.

- Ein zweiter Fall ist ein AIS, das für die **Bewertungen von "Lernergebnissen"** bestimmt ist. Es geht also insbesondere um die Prüfungsbewertung. Die Formulierung des Gesetzes geht etwas allerdings weiter, die Bewertung von Lernergebnissen scheint für sich genommen zu genügen. Erfasst wäre damit eigentlich auch etwa die Korrekturfunktion in einem Sprachlernprogramm, wenn diese ein AIS einsetzt, auch wenn es um nichts weiter geht, als ein Level zu bestehen.

- Der dritte Fall überschneidet sich mit dem ersten: Es geht um AIS, die zur **Bewertung des Bildungsniveaus** dienen soll, das jemand erhalten soll oder zu dem sie zugelassen wird.

Dabei dürften Eignungstests im Vordergrund stehen. Es muss aber um die Bildung gehen – Talent Management mit einer AI-gestützten Bewertung der Eignung für eine andere Stelle wäre hier nicht erfasst (würde aber unter einen anderen Use Case fallen, s. unten).

– Der vierte Fall betrifft AIS, die bestimmungsgemäss zur **Prüfungsaufsicht** in der allgemeinen und beruflichen Bildung zum Einsatz kommen.

Offen ist, wie weit der **Begriff der Bildung** zu verstehen ist. Nach ErwG 56 umfasst dies "Bildungs- und Berufsbildungseinrichtungen oder -programmen auf allen Ebenen", also die schulische Bildung, aber wohl auch die Früh- und Weiterbildung. Kaum erfasst sind dagegen interne Schulungen, die nicht der Weiterbildung dienen, bspw. Compliance-Schulungen. Eine AI-gestützte Auswertung von Testfragen bei einer solchen Schulung sollte daher nicht genügen. Es ist aber ein Grenzfall, und hier werden häufig die arbeitsplatzbezogenen Use Cases (s. unten) greifen (insbesondere die Verhaltens- und Leistungsbewertung eines Mitarbeitenden).

Anhang III Ziff. 4 betrifft spezifisch den **Arbeitsbereich**. Dabei ist zwischen dem Rekrutierungsverfahren und dem Arbeitsverhältnis zu unterscheiden:

– HRAIS sind AIS, die für die **Einstellung oder Auswahl von Bewerbenden** eingesetzt werden sollen. Das ist eine breite Umschreibung, weil weder die Art noch die Wirksamkeit des Einsatzes beschränkt werden. Dieser Use Case ist auch breit gemeint, wie der Text verdeutlicht: Es genügt, wenn ein AIS Bewerbungen "sichtet" oder "filtert".

  Wie bei allen Use Cases von Anhang III muss dieser Einsatz aber in der bestimmungsgemässen Verwendung liegen. Die Formulierung einer Stellenanzeige mit ChatGPT genügt deshalb nicht. Wer dagegen ein AIS baut, das auf der Basis eines OpenAI-Modells Bewerbungen kategorisiert, betreibt ein HRAIS. Es dürfte auch genügen, wenn ein AIS Bewerbungen prüft, wie gut sie auf eine Stellenausschreibung passen – eine Form der semantischen Suche, die einem "Sichten" von Bewerbungen entspricht.

– **Entscheidungen über Arbeitsbedingungen, Beförderung und Kündigung**. Hier ist prima vista unklar, ob das AIS diese Entscheidungen treffen oder nur beeinflussen muss. Aus dem Gesetzestext folgt letzteres: Es geht um den Einsatz für Entscheidungen, die dann – als Folge – Arbeitsbedingungen usw. beeinflussen. Das AIS muss das AIS die Entscheidung aber nicht selbst fällen; es genügt, wenn es dazu bestimmt ist, eine menschliche Entscheidung über solche Punkte zu unterstützen (der englische Text ist klarer: "intended *to be used to* make decisions", nicht "intended to make decisions"). Auch der Gesetzeszweck nach ErwG 57 spricht für diese Auslegung (Schutz der Karriereaussichten und Lebensgrundlagen vor einer "spürbaren Beeinflussung").

– Zwei weitere Use Cases gelten ergänzend. Der eine ist eine **Zuweisung von Aufgaben** auf der Grundlage des Verhaltens oder persönlicher Merkmale oder Eigenschaften, der andere die **Beobachtung und Bewertung der Leistung und des Verhaltens**. Zum HRAIS wird ein AIS damit schon dann, wenn Verhalten ausgewertet wird, auch wenn im Anschluss keine Entscheidung über das berufliche Fortkommen getroffen, vorbereitet oder beeinflusst wird (auch wenn AI-gestützte Leistungs- oder Verhaltensbewertungen i.d.R. auf solche Entscheidungen angelegt sein dürften).

  Ein HRAIS wäre also eine AI-gestützte Auswertung der Performance eines Mitarbeiters in einem Call Center. Kein HRAIS wäre demgegenüber eine AI-gestützte Optimierung von Fahrtrouten des Aussendienstes. Dabei wird zwar grundsätzlich ein Verhalten der entsprechenden Mitarbeitenden ausgewertet. Die Auswertung bezieht sich aber nicht auf dieses Verhalten, sondern abstrahiert davon. Da in einem solchen Fall das berufliche Fortkommen nicht gefährdet ist, sollte er kein HRAIS darstellen. Wird anschliessend aber AI-gestützt bewertet, ob ein Fahrer der optimalen Route folgt, wäre dies ein HRAIS. Dabei dürfte eine menschliche Kenntnisnahme des Ergebnisses keine Voraussetzung sein. Ein Fahrassistent, der abhängig von der tatsächlich befahrenen Route Vorschläge macht, wäre daher wohl ein HRAIS. Dasselbe gilt analog für ein AIS, das im Rahmen der Produktion zur Optimierung der Abläufe eingesetzt wird.

Was alles zum "Arbeitsbereich" gehört, ist damit nicht gesagt. Auch die **selbständige Arbeit** kann erfasst sein, zumal ErwG 57 auch den "Zugang zur Selbstständigkeit" nennt. Alle der genannten Use Cases können dabei auch dann zur Anwendung kommen, wenn die Auswahl, Entscheidung oder Beobachtung bzw. Bewertung nicht einen unselbständigen Mitarbeitenden, sondern einen selbständig Erwerbenden betrifft. Ein arbeitnehmerähnlicher Status, d.h. eine gewisse Abhängigkeit und Subordination, muss aber verlangt werden; andernfalls besteht kein entsprechender Schutzbedarf.

31 Welche Fälle sind bei kritischen Infrastrukturen hochriskant?

Hier sieht der AIA nur einen Fall vor: Ein AIS wird ist (bestimmungsgemäss) ein Sicherheitsbauteil verwendet, das in der Steuerung oder beim Betrieb zum Einsatz kommt, und zwar einer kritischen digitalen Infrastruktur nach Ziff. 8 des Anhangs zur Richtlinie 2022/2557 über die Resilienz kritischer Einrichtungen (CER-Richtlinie, https://dtn.re/D2CV56) und im Bereich des Strassenverkehrs oder bei der Wasser-, Gas-, Wärme- oder Stromversorgung.

32 Welche weiteren Fälle im Privatbereich sind hochriskant?

Anhang III Ziff. 5 regelt drei weitere Fälle, die im Privatbereich relevant sind.

- Der erste betrifft AIS für die "**Kreditwürdigkeitsprüfung und Bonitätsbewertung**" natürlicher Personen (nicht aber juristischer Personen). Das ist relativ breit, weil der AIA nicht definiert, was unter diese Begriffe fällt. Es geht jedenfalls nicht nur um Wirtschaftsauskunfteien und vergleichbare Anbieter von Bonitätsinformationen, sondern auch um Unternehmen, die für sich selbst oder für Gruppengesellschaften entsprechende Bewertungen vornehmen (sofern sie AI-gestützt sind).
- Ausgenommen sind aber AIS, die zur "**Aufdeckung von Finanzbetrug**" verwendet werden. Der Text spricht hier von "verwendet werden" und nicht von "dazu bestimmt sind". Das könnte darauf schliessen lassen, dass ein AIS auch dann kein HRAIS (mehr) ist, soweit sein primärer Zweck zwar in der Bonitätsbeurteilung besteht, es aber nur zur Betrugserkennung eingesetzt wird. Dies widerspricht aber ErwG 58, der insofern enger ist: Ausgenommen sind danach nur AIS, die zur Betrugsprävention "vorgesehen" sind. Er ist gleichzeitig aber auch weiter: AIS, die nach Unionsrecht zur Aufdeckung von Finanzbetrug oder zur Berechnung der Eigenkapitalanforderungen vorgesehen sind, sind kein HRAIS. Das könnte ein Problem sein für einen schweizerischen Finanzdienstleister, der ein AIS zur Berechnung der schweizerischen Kapitalanforderungen einsetzt (also nicht auf Basis des EU-Rechts) und das Ergebnis seinem EU-Stammhaus zur Verfügung stellt und deshalb örtlich unter den AI Act fällt (→ 18).
- Zum HRAIS wird ein AIS ferner dann, wenn es im **Versicherungsbereich** zur Risikoprüfung oder Prämienfestsetzung dient, allerdings nur im Bereich der Lebens- oder Krankenversicherung.
- Ebenfalls ein HRAIS ist ein AIS, das zur Triage von **Notrufen** oder den Einsatz von Sanität, Polizei oder Feuerwehr oder die Priorisierung der ersten Hilfe dient.
- Nach Anhang III Ziff. 6 sind schliesslich auch AIS als HRAIS erfasst, die für die Sachverhaltsermittlung und Rechtsanwendung durch **Schiedsgerichte und Mediatoren** bestimmt sind (neben den staatlichen Gerichten → 33). Erfasst wären bspw. AIS, die aus Akten den Sachverhalt herstellen, wie immer aber unter Vorbehalt einer untergeordneten Unterstützung i.S.v. Art. 6 Abs. 3 (→ 34), bspw. "die Anonymisierung oder Pseudonymisierung gerichtlicher Urteile, Dokumente oder Daten, die Kommunikation zwischen dem Personal oder

Verwaltungsaufgaben" (ErwG 61). Auch im Zusammenhang mit der Wahl- und Abstimmungsbeeinflussung können AIS im Privatbereich ein HRAIS sein (→ 33).

33 Welche Fälle sind im öffentlichen Bereich hochriskant?

Anhang III enthält einige Use Cases, die nur im öffentlichen Bereich relevant sind (aber einschliesslich von Unternehmen, die im Auftrag einer Behörde tätig sind).

Anhang III Ziff. 5 betrifft AIS zur Verwendung durch oder für Behörden für die Beurteilung, ob ein Anspruch auf "**grundlegende öffentliche Unterstützungsleistungen und -dienste**" besteht, einzuschränken oder aufzuheben ist. Das betrifft bspw. die Sozialversicherung oder soziale Hilfe. Diese Fälle sind aber jeweils auf die Anwendung gegenüber natürlichen Personen beschränkt.

Anhang III Ziff. 6 betrifft verschiedene Use Cases im Bereich der **Strafverfolgung**, und Ziff. 7 im Bereich der **Migration, des Asylwesens und der Grenzkontrolle**. Ziff. 8 lit. a betrifft sodann AIS zur Verwendung durch oder für eine Justizbehörde (einschliesslich der privaten Streitbeilegung → 32) als Unterstützung bei der Sachverhaltsermittlung und Rechtsanwendung. Nach lit. b sind AIS ferner auch dann hochriskant, wenn sie dazu dienen, das Ergebnis einer Wahl oder Abstimmung oder das Wahlverhalten zu beeinflussen. Es muss allerdings um eine direkte Beeinflussung gehen – nicht erfasst sind AIS Instrumente zur administrativen Unterstützung von Kampagnen.

34 Gibt es Fälle, bei denen ein HRAIS ausnahmsweise nicht als hochriskant gilt?

Ja. Anders als bei den produktbezogenen Hochrisikofällen ist es bei den einsatzbezogenen Einstufungen nach Anhang III (→ 28) möglich, im Sinne einer Ausnahme das fehlende hohe Risiko zu belegen.

Das trifft nach Art. 6 Abs. 3 unter zwei kumulativen Voraussetzungen zu:

- Erstens muss der **Verwendungszweck des AIS harmlos** sein, weil es weder ein grösseres Risiko bedeutet noch eine Entscheidung wesentlich beeinflusst (ErwG 53). Das ist der Fall, wenn es nur dazu bestimmt ist,
  - eine "eng gefasste Verfahrensaufgabe" durchzuführen (bspw. unstrukturierte Daten zu strukturieren oder Daten zu kategorisieren), oder
  - als blosser zusätzlicher Layer das Ergebnis einer menschlichen Tätigkeit zu verbessern (bspw. beim inzwischen üblichen "Verbessern" eines von einem Menschen verfassten Textes), oder Entscheidungsmuster oder Abweichungen von früheren Entscheidungsmustern zu erkennen (bspw. bei einer Überprüfung, ob eine menschliche Prüfung von einem vorgegebenen Muster abweicht), oder
  - eine bloss vorbereitende Aufgabe für eine Bewertung durchzuführen (bspw. bei einer Übersetzung von Texten für die weitere menschliche Verwendung); jeweils näher in Art. 6 Abs. 3 und ErwG 53). Die EU-Kommission (→ 51) sollte hier Konkretisierungen vorschlagen.
- Zweitens darf das AIS **kein Profiling** vornehmen (a.a.O.). Für den Begriff verweist der AIA auf die DSGVO (https://dtn.re/8YoXjh; Art. 4 Nr. 4).

Ein Anbieter, der diese Ausnahme in Anspruch nehmen will, muss diese Einschätzung vor dem Inverkehrbringen bzw. der Inbetriebnahme dokumentieren (Art. 6 Abs. 4). Er muss das AIS ferner gleich wie ein HRAIS registrieren (Art. 49).

Kernpflichten bei HRAIS

35 Welches sind die wesentlichen Pflichten entlang der Wertschöpfungskette?

Nicht alle Zwischenschritte im Rahmen der Wertschöpfungskette sind Hauptauslöser von Pflichten und Auflagen. Grundsätzlich müssen AIS zum Zeitpunkt des Inverkehrbringens bzw. der Inbetriebnahme die grundlegenden Anforderungen erfüllen (→ 15). Aus einer praktischen Sicht lösen aber auch andere Vorgänge bestimmte Pflichten aus.

Diese Pflichten lassen sich übersichtsweise wie folgt aufschlüsseln, wobei die Zuordnung zu einzelnen Phasen nicht trennscharf sein kann, weil der AIA nicht alle dieser faktisch zu unterscheidenden Stufen rechtlich als Anknüpfungspunkt von Pflichten bestimmt. Einzelheiten zu den einzelnen Pflichten finden sich in den verwiesenen Fragen bzw. Antworten. Zu beachten ist weiter, dass bei den Produkten nach Anhang III Abschnitt B nicht der AIA gilt, sondern die in die jeweilige Produktregulierung übernommenen Pflichten (→ 28).

	System	Rolle	Auslöser	Rechtliche Folgen und Anforderungen
Anbieter				
1	**HRAIS**	**Anbieter**	**Beschaffung von Systemkomponenten**	Bezieht der Anbieter Komponenten von einem Zulieferer – es wird um Softwarekomponenten gehen, weil in einer Kombination von Hard- und Software lediglich die Software als AIS gelten dürfte und der Bezug von Hardware deshalb keinen Einbau in ein AIS darstellt (→ 28) –, muss er mit dem Lieferanten einen **Vertrag schliessen**. Der Vertrag muss schriftlich sein, d.h. wohl in Textform dokumentiert, und die für den Anbieter des HRAIS wesentlichen Punkte regeln (Art. 25 Abs. 4). Das AI Office (→ 52) soll hier Vorlagen liefern. Ausgenommen von dieser Pflicht ist die Lieferung einer Nicht-GPAI im Rahmen einer freien und quelloffenen Lizenz (FOSS), doch werden entsprechende Softwareanbieter ermutigt, für HRAIS-Anbieter relevante Informationen bereitzustellen (ErwG 89).
2	**HRAIS**	**Anbieter**	**Training**	Ein HRAIS muss nicht zwingend trainiert werden – das ist nicht eine eigene Pflicht (Art. 10 Abs. 6; auch nicht als Risikomitigierungsmassnahme: ErwG 65), sondern vielmehr ein Umstand, der zur Einstufung als AIS führen kann (→ 13). Im Falle eines Trainings greifen aber bestimmte Anforderungen. Zunächst stellt sich die Frage, mit **welchen Daten** ein HRAIS trainiert werden soll oder darf. Dafür gibt Art. 10 Abs. 3 (Data Governance) Anforderungen vor: Testdaten müssen Merkmale oder Elemente berücksichtigen, die für die Rahmenbedingungen des HRAIS in seiner bestimmungsgemässen Verwendung typisch sind, also aussagekräftig. Das kann die Verwendung von **Personendaten** oder sogar besonders schützenswerter Personendaten (→ 58) voraussetzen, bspw. bei Systemen, die Bewerbungen einstufen und so trainiert werden müssen, dass sie einen möglichst schwachen Bias in Bezug auf Alter, Geschlecht, ethnischem Hintergrund usw. aufweisen (manche Hersteller legen ihrer Kundendokumentation deshalb Bias-Audits bei). Art. 10 Abs. 5 enthält deshalb eine Rechtsgrundlage für die Verwendung solcher Daten zu Test- und Trainingszwecken, unter den Voraussetzungen von Abs. 5 lit. a-f. Für das Training selbst muss der Anbieter sodann einige Entscheidungen treffen und **dokumentieren**. Dies legt Art. 10 Abs. 2 fest. Es geht vor allem um die Beschaffung von Trainingsdaten, die Vorbereitung von Testdaten (z.B. Labelling, Tagging usw.), die Definition von Annahmen und Zielgrössen, die

	System	Rolle	Auslöser	Rechtliche Folgen und Anforderungen
				Metrik zur Messung, ob Ziele erreicht bzw. Annahmen zutreffend sind, oder um die Vermeidung von Verzerrungen (Bias). Auch in der Trainingsphase ist das genannte **Risikomanagementsystem** (RMS) für HRAIS relevant (vgl. Art. 9). Zwar muss der Anbieter wie erwähnt kein Training als Risikomitigierungsmassnahme durchführen, wird dazu aber dennoch eingeladen (ErwG 65). Insofern sollte das RMS selbstverständlich auch die Trainingsphase abdecken.
3	**HRAIS**	**Anbieter**	**Testen**	Anders als ein Training ist ein Test eine eigene **Pflicht** des HRAIS-Anbieters (Art. 9 Abs. 6). HRAIS müssen getestet werden, damit das Risiko ermittelt und ggf. mitigiert werden kann. Tests müssen zum geeigneten Zeitpunkt, aber vor dem Inverkehrbringen oder der Inbetriebnahme durchgeführt werden (Abs. 8). Zu Tests durch Anbieter von GPAI-Modellen mit systemischen Risiken siehe Q41. Für die Durchführung der Tests finden sich in **Art. 9 und 10** Vorgaben. Die Anforderungen an Trainingsdaten gelten auch für Testdaten (Art. 9 Abs. 6; das betrifft auch eine etwaige Verwendung von Personendaten). Nach Art. 9 Abs. 7 können Tests unter Umständen für höchstens 12 Monate auch unter **Realbedingungen** erfolgen – sofern die Voraussetzungen von Art. 60 AIA eingehalten werden. Solche Tests erfordern u.a. einen eigenen Plan, der von der zuständigen Marktüberwachungsbehörde zu genehmigen ist (Art. 60 Abs. 4 lit. a und b).
4	**HRAIS**	**Anbieter**	**Inverkehrbringen bzw. Inbetriebnahme**	Rechtlich ist der Zeitpunkt des Inverkehrbringens bzw. der Inbetriebnahme des HRAIS massgebend für die **meisten Pflichten** des Anbieters. Er muss diese daher bei der Planung und beim Design eines AIS, das potentiell ein HRAIS ist, berücksichtigen. Zunächst muss der Anbieter die **technische Dokumentation** erstellen (Art. 11 und Anhang IV). Das ist das Herzstück: Die technische Dokumentation dient dazu, die Einhaltung der grundlegenden Anforderungen zu dokumentieren, und sie ist entsprechend auch Grundlage der Konformitätsbewertung. Sie enthält insbesondere eine Beschreibung des HRAIS, seiner Bestandteile, seiner Entwicklung bzw. seines Trainings einschliesslich der dafür verwendeten Daten und der Validierung und der massgeblichen Tests, seiner Funktionsweise und Architektur, der Gewährleistung der menschlichen Aufsicht (→ 37), seiner Kontrolle, des Risikomanagement-Systems und des Verfahrens der Marktüberwachung (Anhang IV).

System	Rolle	Auslöser	Rechtliche Folgen und Anforderungen
			Auch die **Betriebsanleitung** (Art. 3 Nr. 15 und Art. 13 Abs. 3) gehört zur technischen Dokumentation (Anhang IV Ziff. 1 lit. h). Diese nennt und definiert die bestimmungsgemässe Verwendung des HRAIS (Art. 3 Nr. 15), die darüber bestimmt, ob das AIS ein HRAIS nach Anhang III ist (→ 32) und dazu beiträgt, den Verantwortungsbereich des Anbieters abzugrenzen, und die ein wesentlicher Massstab für die Compliance-Anforderungen ist (vgl. etwa Art. 8 Abs. 1, Art. 10 Abs. 3 oder Art. 26 Abs. 6 AIA). Die Betriebsanleitung muss präzise, vollständige, korrekte, eindeutige und verständliche Informationen enthalten und digital oder physisch, aber barrierefrei bereitgestellt werden (Art. 13 Abs. 3) und mindestens die Informationen nach Art. 13 Abs. 3 lit. a-f enthalten. Dazu gehören u.a. die Zweckbestimmung, die Eigenschaften und die Leistungsgrenzen des HRAIS, die Massnahmen zur Gewährleistung menschlicher Aufsicht, die Lebensdauer des HRAIS und Angaben zur Pflege und zu Updates und eine Beschreibung der Logfähigkeit. Auf Basis der technischen Dokumentation muss der Anbieter sodann rechtzeitig vor dem Inverkehrbringen bzw. der Inbetriebnahme das **Konformitätsbewertungsverfahren** durchlaufen (→ 15), und er muss für jedes HRAIS eine EU-**Konformitätserklärung** ausstellen und zuhanden der Behörden aufbewahren (Art. 47). Zudem muss er jeweils ein physisches oder digitales **CE-Zeichen** anbringen, abgesehen von der Angabe seines Namens bzw. seiner Marke und einer Kontaktadresse (Art. 16 lit. b → 15). Zu den wesentlichen Pflichten gehören ferner die folgenden: – **QMS**: Nach Art. 17 muss der Anbieter über ein Qualitätsmanagementsystem (QMS) verfügen, das allgemein die Einhaltung des AIA "gewährleistet", also ein System aus Policies, Prozessen und Anweisungen, das alle Phasen des HRAIS abdeckt, einschliesslich eines Compliance-Konzepts mit Zuständigkeiten und Verantwortlichkeiten, Angaben zur Entwicklung und zum Testen und Validieren des HRAIS, ggf. die harmonisierten Normen, die für die Konformitätsbewertung zur Anwendung kommen, die Data Governance (→ 36), die Marktüberwachung (→ 43), den Umgang mit Incidents (→ 45), die Kommunikation mit Behörden, die erforderliche Dokumentation und das Ressourcenmanagement. Auch das Risikomanagementsystem ist Teil des QMS (Art. 17 Abs. 1 lit. g; das RMS kann separat geführt werden, muss im QMS aber abgedeckt werden). – **RMS**: Der Anbieter muss für jedes HRAIS ein Risikomanagementsystem (RMS) einrichten, anwenden, dokumentieren und aufrechterhalten (Art. 9). Das RMS muss das HRAIS durch seinen Lebenszyklus hindurch – auch nach dem Inverkehrbringen bzw. der Inbetriebnahme – begleiten und aktuell gehalten werden – das verlangt eine entsprechende Governance. Insbesondere müssen

System	Rolle	Auslöser	Rechtliche Folgen und Anforderungen
			Risiken für die Gesundheit, die Sicherheit oder die Grundrechte insbesondere auch schutzbedürftiger Personen laufend erkannt und bewertet werden, und zwar nicht nur in Bezug auf den bestimmungsgemässen Einsatz, sondern auch den vorhersehbaren Missbrauch (Art. 9 Abs. 2 lit. b), und sie müssen schon bei der Design- und Entwicklungsphase angemessen mitigiert werden, soweit der Anbieter die Risiken mitigieren kann (lit. d). Dazu gehört auch bspw. die Information des Betreibers oder seine Schulung (Art. 9 Abs. 5 lit. c). Die identifizierten und akzeptierten Risiken sollten dann in der Betriebsanleitung genannt werden. Der Anbieter kann sich beim RMS an entsprechenden Standards orientieren (→ 61). – **Logfähigkeit sicherstellen**: Der Anbieter muss sicherstellen, dass das System technisch logfähig ist (Art. 12). Was geloggt werden muss, gibt Art. 12 Abs. 2-3 vor. – **Verständlichkeit des Output** (Art. 13): Der Anbieter muss sicherstellen, dass der Output des Systems für den Betreiber klar und verständlich ist. Dazu dient die Betriebsanleitung (Art. 3 Nr. 15), aber auch Designmassnahmen werden erforderlich sein. – **Menschliche Aufsicht** (Art. 14): Das HRAIS muss so gestaltet sein, dass es eine wirksame menschliche Aufsicht ermöglicht. Dazu kommen ins HRAIS eingebaute Massnahmen in Betracht (z.B. Benutzerschnittstellen, ein Kill Switch usw.), aber auch Anleitungen zuhanden des Betreibers, die ihm ein ausreichendes Verständnis des HRAIS ermöglichen (vgl. Art. 11, Art. 14 Abs. 4 und Anhang IV). – **Zuverlässigkeit, Robustheit und Cybersicherheit** (Art. 15): Ein HRAIS muss so konzipiert sein, dass es zuverlässig und robust ist und ein ausreichendes Mass an Cybersicherheit gewährleistet. Der Anbieter muss deshalb u.a. sicherstellen, dass das HRAIS gegen physische und digitale Bedrohungen ausreichend resistent ist und geeignete Massnahmen zum Schutz der Integrität, Vertraulichkeit und Verfügbarkeit des HRAIS greifen. Bei Systemen, die auch nach dem Inverkehrbringen bzw. der Inbetriebnahme lernen, muss das Risiko eines Bias und von Feedback-Loops mitigiert werden. Die EU-Kommission (→ 51) soll zur Entwicklung von Benchmarks und Messgrössen beitragen (Art. 15 Abs. 2 AIA). – **Accessibility by Design** (Art. 16): Die Accessibility muss in das Design des HRAIS integriert werden. Die Anforderungen ergeben sich im Einzelnen aus der RL 2016/2102 über den barrierefreien Zugang zu den Websites und mobilen Anwendungen öffentlicher Stellen und der RL 2019/882 über die Barrierefreiheitsanforderungen für Produkte und Dienstleistungen (Art. 16 lit. l).

	System	Rolle	Auslöser	Rechtliche Folgen und Anforderungen
				– **Registrierung**: Anbieter müssen HRAIS bei der Kommission (→ 51) registrieren, wenn sie nach Anhang III (Use Cases) als HRAIS einzustufen sind (→ 28). Dazu müssen sie mindestens die Informationen nach Anhang VIII Abschnitt A bereitstellen.
5	**HRAIS**	**Anbieter**	**Auftreten besonderer Risiken**	Erlangt der Anbieter Kenntnis von besonderen Risiken i.S.v. Art. 79 Abs. 1, muss er sofort die Ursachen untersuchen und die zuständigen Marktüberwachungsbehörden informieren (Art. 82 Abs. 2 → 45).
6	**HRAIS**	**Anbieter**	**Eintritt eines schwerwiegenden Vorfalls**	Bei Feststellung eines schwerwiegenden Vorfalls (→ 45) muss der Anbieter sofort die zuständigen Marktüberwachungsbehörden (→ 55) informieren, den Vorfall untersuchen und Risiken einschätzen mitigieren. Für Anbieter von GPAIM mit systemischen Risiken s. unten.
7	**AIS**	**Anbieter**	**Inverkehrbringen oder Inbetriebnahme in der EU**	Im Fall des Inverkehrbringens oder der Inbetriebnahme eines AIS in der EU fällt der Anbieter unter den AIA (→ 18) und muss er einen Bevollmächtigten in der EU bestellen (→ 26).
8	**AIS**	**Anbieter**	**Verwendung von Output in der EU**	Auch wenn eine Stelle ein AIS so verwendet, dass sein Output bestimmungsgemäss in der EU verwendet wird, fällt er in den **Anwendungsbereich** des AIA (→ 18) und muss er einen **Bevollmächtigten** in der EU bestellen (→ 26).
9	**AIS**	**Anbieter**	**Umgang mit AIS**	Der Umgang eines Anbieters mit AIS löst auch für ihn die Anforderung an AI Literacy aus (→).
10	**AIS**	**Anbieter**	**Generative AIS**	Bei AIS – das werden vor allem GPAIS sein, erfasst werden aber auch andere AIS –, die synthetische Inhalte (Audio, Bild, Video, Text) generieren, müssen Anbieter dafür sorgen, dass der Output in einem **maschinenlesbaren Format gekennzeichnet** ist, damit er als künstlich erzeugt oder manipuliert erkennbar ist (“Watermarking” → 37).
11	**AIS**	**Anbieter**	**AIS zur direkten Interaktion mit Betroffenen**	Ist ein AIS (ggf. einschliesslich eines HRAIS) für die direkte Interaktion mit betroffenen Personen bestimmt, muss der Anbieter schon in der Konzeptions- und Entwicklungsphase dafür sorgen, dass die natürlichen Personen über die **Interaktion mit einem AIS informiert** werden (wenn es in den gegebenen Umständen nicht offensichtlich ist → 37).

	System	Rolle	Auslöser	Rechtliche Folgen und Anforderungen
Produkthersteller				
12	**HRAIS**	**Produkthersteller**	**Einbau eines AIS in ein Produkt**	Hersteller eines regulierten Produkts, das unter eine Produktregulierung nach Anhang I fällt, weil ein AIS als Sicherheitsbauteil (i.S.v. Art. 3 Nr. 14) verbaut wurde, und die das Produkt im eigenen Namen in Verkehr bringen oder in Betrieb nehmen, gelten als Anbieter i.S.d. AIA (Art. 25 Abs. 3) und haben die entsprechenden Pflichten.
Einführer und Händler				
13	**HRAIS**	**Einführer**	**Import**	Die Pflichten des Einführers (“Importer” → 23) sind wesentlich enger als jene des Anbieters, weil die Hauptverantwortung beim Anbieter bleibt. Der Einführer hat zunächst vor allem die Pflicht, Compliance-Massnahmen des Anbieters nachzuprüfen, und zweifelt er an der Compliance des HRAIS, darf er es HRAIS nicht in Verkehr bringen. Falls er auf Risiken i.S.v. Art. 79 Abs. 1 stösst, hat er zudem den Anbieter, die Bevollmächtigten und die Marktüberwachungsbehörden zu informieren (Art. 23 Abs. 2 → 45). Weitere Pflichten ergeben sich aus Art. 23 Abs. 3-7.
14	**HRAIS**	**Betreiber, Einführer, Händler**	**Auftreten besonderer Risiken**	Hat ein Betreiber oder ein Einführer Grund zur Annahme, dass ein HRAIS mit besonderen Risiken für die Gesundheit, die Sicherheit oder Grundrechte verbunden ist (Art. 79), muss er sofort sowohl den Anbieter oder Händler (im Fall des Betreibers) bzw. den Anbieter und dessen Bevollmächtigen (im Fall des Einführers) bzw. den Anbieter und den Einführer oder jede andere beteiligte Stelle (im Fall des Händlers) als auch die zuständige Marktüberwachungsbehörde informieren und den Einsatz des HRAIS aussetzen (Art. 26 Abs. 5, Art. 23 Abs. 2 und Art. 24 Abs. 4; Art. 82 Abs. 2 → 45).
15	**HRAIS**	**Händler**	**Vertrieb**	Händler ist, wer ein HRAIS auf Markt bereitstellt (→ 20). Die Pflichten der Händler sind ähnlich wie jene des Einführers (Art. 24).
Betreiber				
16	**HRAIS**	**Betreiber**	**Einsatz**	Betreiber (→ 21) müssen die eingesetzten HRAIS **inventarisieren** (das folgt indirekt aus Art. 26). Sie müssen zudem sicherstellen, dass alle relevanten Betriebsdaten automatisch **protokolliert** und für

	System	Rolle	Auslöser	Rechtliche Folgen und Anforderungen
				einen festgelegten Zeitraum aufbewahrt werden, und sie müssen sich beim Einsatz des HRAIS an die **Anweisungen des Anbieters** halten (Art. 26 Abs. 1). Sie müssen ferner dafür sorgen, die Inputdaten zweckkonform (d.h. dem Zweck des HRAIS angemessen) und ausreichend repräsentativ sind (Art. 26 Abs. 4 → 36). Zentral ist zudem die **menschliche Überwachung**: Der Betreiber muss sicherstellen, dass während des Betriebs eine menschliche Überwachung möglich ist (Art. 26 Abs. 2), und er muss den Betrieb des Systems kontinuierlich überwachen (Art. 26 Abs. 5). Bei Verdacht auf ein besonderes Risiko i.S.v. Art. 79 Abs. 1 (→ 45) müssen sie den Anbieter oder Händler und die Marktüberwachungsbehörde entsprechend informieren und die Verwendung des HRAIS einstellen (Art. 26 Abs. 5; was voraussetzt, dass sie entsprechend reagieren können). Im Fall eines schwerwiegenden Vorfalls (→ 45) festgestellt, ist sofort den Anbieter und dann der Einführer oder Händler und die Marktüberwachungsbehörde zu informieren (vgl. auch Art. 73).
17	**HRAIS**	**Betreiber**	**Einsatz am Arbeitsplatz**	Setzt ein Arbeitgeber ein HRAIS am Arbeitsplatz ein, muss er die Arbeitnehmer und die Arbeitnehmervertreter informieren, dass sie von der Verwendung betroffen sein werden (Art. 26 Abs. 7). Vorbehalten sind Mitwirkungspflichten nach dem anwendbaren Recht.
18	**HRAIS**	**Betreiber**	**Verwendung für Entscheidungen**	Besondere Anforderungen gelten, wenn ein HRAIS für Entscheidungen eingesetzt werden soll (es kann auch sein, dass ein AIS dadurch zum HRAIS wird: Art. 25 und Anhang III → 28). Wenn das HRAIS **Entscheidungen trifft**, die rechtliche oder andere signifikante Auswirkungen haben, müssen die Betroffenen informiert werden (Art. 13 und 25 Abs. 11 → 37), und bei **automatisierten AI-Entscheidungen** haben betroffene Personen das Recht, dagegen Einspruch zu erheben (Art. 86; zudem können natürlich die einschlägigen Anforderungen des anwendbaren Datenschutzrechts greifen). Zudem muss der Betreiber sicherstellen, dass die Eingabedaten für das System relevant, korrekt und aktuell sind (s. oben).
19	**HRAIS**	**Betreiber**	**Biometrische Fernidentifikation**	Wird ein HRAIS für die biometrische Fernidentifikation 4i.S.v. Anhang III Ziff. I lit. a eingesetzt, müssen die Ergebnisse durch mindestens zwei natürliche, kompetente Personen getrennt überprüft und bestätigt werden, bevor Entscheidungen gefällt oder Massnahmen getroffen werden (Art. 15 Abs. 4).

	System	Rolle	Auslöser	Rechtliche Folgen und Anforderungen
20	**HRAIS**	**Betreiber**	**Einsatz eines Emotionserkennungssystems oder zur biometrischen Kategorisierung**	Beim Einsatz eines Emotionserkennungssystems oder eines Systems zur biometrischen Kategorisierung müssen die Betreiber die betroffenen Personen über den Betrieb und die verwendeten Personendaten informieren (→ 37).
21	**HRAIS**	**Betreiber**	**Eintritt eines schwerwiegenden Vorfalls**	Bei Feststellung eines schwerwiegenden Vorfalls muss der Betreiber sofort sowohl den Anbieter und dann den Einführer oder Händler als auch die zuständigen Marktüberwachungsbehörden informieren (Art. 26 Abs. 5 und Art. 72 → 45).
22	**AIS**	**Betreiber**	**Betrieb**	Im Betrieb eines AIS gelten nur die Anforderung an AI Literacy (→ 38).
23	**AIS**	**Betreiber**	**Einsatz für Deepfakes**	Wird ein AIS (das auch ein HRAIS sein kann) für Deepfakes verwendet, muss der Betreiber die künstliche Herstellung offenlegen (Art. 50 Abs. 4 → 37).
24	**AIS**	**Betreiber**	**Generierung von Output**	Setzen Betreiber ein AIS ein, um Text zu erzeugen oder manipulieren, und wird der Text veröffentlicht, um die Öffentlichkeit über Angelegenheiten von öffentlichem Interesse zu informieren, müssen sie die künstliche Herstellung bzw. die Manipulation offenlegen (Art. 50 Abs. 4 → 37).
GPAIM				
25		**Anbieter**	**Inverkehrbringen eines GPAIM in der EU**	Anbieter eines GPAIM fallen in den Anwendungsbereich des AIA, wenn sie in der EU ein GPAIM in Verkehr bringen (→ 18). In diesem Fall müssen sie einen Bevollmächtigten in der EU bestellen (→ 26).
26		**Anbieter**	**Anbieten eines GPAIM für den Einbau in ein AIS**	Der AIA versteht GPAIM nicht als HRAIS, sondern als Vorstufe zu einem AIS (→ 39). Der Anbieter des GPAI-Modells muss **Anbieter, die GPAIM verbauen**, deshalb über das GPAI-Modell und seine Entwicklung informieren, nach den Vorgaben von Anhang XII (Art. 53 Abs. 1 lit. b). Er muss insbesondere eine technische Dokumentation erstellen (Art. 53 Abs. 1 lit. a), aber nicht wie bei HRAIS-Anbietern nach Anhang IV, sondern nach dem eigenen Anhang XI. Weil GPAI-Modelle meist LLMs sind, die mit einer Masse von Daten trainiert werden, muss der Anbieter des Modells ferner eine **Policy zur Einhaltung des europäischen Urheberrechts** haben (Art. 53

	System	Rolle	Auslöser	Rechtliche Folgen und Anforderungen
				Abs. 1 lit. c; siehe Q0), und er muss Angaben über die Trainingsdaten öffentlich zur Verfügung stellen (Art. 53 Abs. 1 lit. d; dafür soll das AI Office → 52 eine Vorlage ausarbeiten). Von diesen Pflichten ausgenommen sind allerdings Anbieter, die das GPAIM im Rahmen einer freien und quelloffenen Lizenz (**FOSS**), wenn sie die Parameter des Modells öffentlich zugänglich machen. Davon gilt wiederum eine Gegenausnahme bei GPAI-Modellen mit systemischen Risiken (→ 39). Die allgemeinen Anforderungen an HRAIS-Anbieter gelten für GPAIM-Anbieter demgegenüber nicht (→ 39; solange sie nicht auch HRAIS-Anbieter sind).
27		**Anbieter**	**Anbieten eines GPAIM mit systemischen Risiken**	Der Anbieter eines GPAIM mit systemischen Risiken (→ 39) hat die Pflichten aller GPAIM-Anbieter. Zusätzlich muss er das Modell zunächst der EU-**Kommission melden**, spätestens zwei Wochen, nachdem das Modell die Schwelle der systemischen Risiken erreicht hat (Art. 52 Abs. 1). Die Kommission pflegt eine entsprechende, öffentliche Liste (Art. 52 Abs. 6), wobei der Anbieter versuchen kann, sein Modell als nicht systemisch relevant streichen zu lassen (→ 41). Weiter ist der entsprechende Anbieter nach Art. 55 Abs. 1 verpflichtet, – systemische Risiken zu bewerten und ggf. zu mitigieren, – das Modell mit Blick auf das Risikomanagement zu bewerten, auch durch Angriffstests (Adversarial Testing bzw. Red Teaming), – Informationen über schwerwiegende Vorfälle (→ 45) und mögliche Minderungsmassnahmen zu dokumentieren und das AI Office (→ 52) sofort entsprechend zu informieren, – angemessene Cybersicherheit zu gewährleisten.

36 Was gilt für das Training, Validieren und Testen von KI-Systemen?

Tests und Validierungen und insbesondere das Training sind zentrale Aspekte bei AIS. Der AIA enthält dafür besondere Regelungen:

- Anbieter sind verpflichtet, das HRAIS vor dem Inverkehrbringen bzw. vor der Inbetriebnahme zu **testen** (Art. 9 Abs. 6).
- Für die für das Training und für Testzwecke verwendeten Daten (vgl. Art. 3 Nr. 29 und 31) müssen geeignete "**Daten-Governance- und Datenverwaltungsverfahren**" zur Anwendung kommen (Art. 10 Abs. 1 und 2). Dabei ist insbesondere zu regeln, wie die entsprechenden konzeptionellen Entscheidungen getroffen werden, welche Daten erforderlich sind und wie sie beschafft werden (insbesondere Personendaten), wie Daten aufbereitet werden (bspw. durch Annotation, Labelling, Bereinigung, Aktualisierung, Anreicherung und Aggregierung), wie Prüfhypothesen gebildet werden und wie mit einem möglichen Bias umzugehen ist (Art. 10 Abs. 2).
- Trainings-, Validierungs- und Testdaten müssen – mit Blick auf die bestimmungsgemässe Verwendung des HRAIS – möglichst **relevant, repräsentativ, fehlerfrei und vollständig** sein. Dazu gehört auch, dass sie über geeignete statistische Merkmale verfügen (Art. 10 Abs. 3), und den Kontext ihrer Verwendung abbilden bzw. berücksichtigen (Abs. 4).
- Unter Umständen kann ein Bias nur verhindert oder entdeckt werden, wenn die für Trainings, Tests und Validierungen verwendeten Daten **Personendaten** enthalten. Für diesen Fall enthält Art. 10 Abs. 5 ausnahmsweise eine Rechtsgrundlage i.S.v. Art. 6 und 9 DSGVO, d.h. sogar für besondere Kategorien personenbezogener Daten, sofern dabei bestimmte Voraussetzungen zur Gewährleistung der Datensparsamkeit und zum Schutz der betreffenden Daten eingehalten werden. Dies muss im Verarbeitungverzeichnis dokumentiert werden.
- Der Anbieter muss die nachgelagerten Akteure informieren, insbesondere über die technische Dokumentation und die Betriebsanleitung (→ 35). Die **technische Dokumentation** muss u.a. Angaben über das Training und die verwendeten Trainingsdatensätze enthalten (Anhang IV Ziff. 2 lit. d; ebenso für Anbieter eines GPAIM nach Anhang XI Ziff. 2 lit. b und nach Anhang XII Ziff. 2 lit. c, wenn das GPAIM in ein AIS integriert werden soll), und die **Betriebsanleitung** muss ebenfalls Angaben über die verwendeten Trainings-, Validierungs- und Testdatensätze enthalten (Art. 13 Abs. 3 lit. b Ziff. 6).
- HRAIS müssen generell ein ausreichendes Mass an **Cybersicherheit** gewährleisten. Dazu gehört auch ein ausreichender Schutz gegen Angriffe in der Trainingsphase, bspw. durch Manipulation der Trainingsdaten ("data poisoning") oder vortrainierter Komponenten wie bspw. einem GPAIM, die beim Training zum Einsatz kommen ("model poisoning"; Art. 15 Abs. 5).
- Anbieter eines GPAIM müssen analog zum Anbieter eines HRAIS in der technischen Dokumentation das Trainings- und Testverfahren **dokumentieren** (Art. 53 Abs. 1 lit. a), und sie müssen eine Zusammenfassung der für das Training verwendeten Inhalte erstellen und **veröffentlichen** (Art. 53 Abs. 1 lit. d; mit Ausnahme von FOSS).
- Die Menge der für das Training verwendeten Berechnungen ist massgebend für die Einstufung eines GPAIM als ein solches mit **systemischen Risiken** (Art. 51 Abs. 2).
- Die **Marktüberwachungsbehörden** können Zugang u.a. zu den Trainings-, Validierungs- und Testdatensätzen verlangen (Art. 74 Abs. 12).
- Erleichterungen für das Training gelten sodann im Rahmen der **Reallabore** (→ 48).

Diese Pflichten richten sich naturgemäss an die Anbieter. **Betreiber** haben andere, eigene Pflichten mit Bezug auf die Datenqualität (→ 35).

37 Wie adressiert der AI Act die Transparenzpflichten für KI-Systeme, insbesondere bei automatisierten Entscheidungen?

Der AI Act legt besonderes Gewicht auf Transparenz, insbesondere bei AIS, die Entscheidungen treffen. Das kann auch bei AIS gelten, die **keine HRAIS** sind. Insbesondere das eigene Kapitel IV enthält mit seinem einzigen Art. 50 entsprechende Vorgaben, wobei die ersten beiden Absätze Anbieter und die folgenden beiden Absätze Betreiber betreffen.

Anbieter haben insbesondere die folgenden Pflichten:

- **Systemgestaltung**: Anbieter müssen das HRAIS so konzipieren, dass der Betrieb transparent ist, d.h. dass die Ausgaben interpretier- und bewusst verwendbar sind (Art. 13 Abs. 1). Wie dies zu gewährleisten ist, gibt der AIA nicht abschliessend vor.

- **Betriebsanleitung**: HRAIS müssen jeweils von einer Betriebsanleitung begleitet werden (→ 35 Nr. 6).

- **Interaktion mit Betroffenen**: Bei AIS, die zur Interaktion mit Betroffenen vorgesehen sind (z.B. Chatbots), müssen diese über die Interaktion mit einem AIS ins Bild gesetzt werden – es sei denn, es sei in den Umständen offensichtlich (Art. 50 Abs. 1), bspw. bei einem Übersetzungsdienst oder einem Chatbot wie ChatGPT. Dies müssen ggf. die Anbieter des entsprechenden AIS sicherstellen. Dazu kann oft schon die Bezeichnung als "Bot" genügen.

- **Synthetische Inhalte**: Anbieter von AIS müssen synthetische Ausgaben in einem maschinenlesbaren Format kennzeichnen und sicherstellen, dass sie als künstlich erzeugt oder manipuliert erkennbar sind (Art. 50 Abs. 2). Diese Pflicht betrifft wiederum Anbieter, nicht Betreiber (dazu den nächsten Punkt). Zu verweisen ist hier auf die Arbeiten der Coalition for Content Provenance and Authenticity (C2PA; https://c2pa.org).

  Ausgenommen sind HRAIS mit einer bloss unterstützenden Funktion für eine Standardbearbeitung oder ohne wesentliche Veränderung des Inputs. Keine Kennzeichnungspflicht gilt daher etwa für DeepL oder ChatGPT bearbeitete, von einem Menschen verfasste Texte. Über den Wortlaut hinaus muss das analog zu Art. 50 Abs. 4 auch gelten, wenn ein Text von einem AIS generiert, aber von einem Menschen überarbeitet oder zumindest relevant kontrolliert wurde; in diesem Fall hat sich der Mensch den Text zu eigen gemacht, weshalb er nicht mehr als synthetisch behandelt werden sollte.

- **Menschliche Aufsicht**: Art. 14 enthält Vorgaben zur Gewährleistung menschlicher Aufsicht, die ebenfalls einen Transparenzaspekt haben.

Betreiber haben insbesondere die folgenden Pflichten:

- **Deepfakes**: Deepfakes sind nach Art. 3 Nr. 60 Bild-, Ton- oder Videoinhalte, die wirklichen Personen, Gegenständen, Orten, Einrichtungen oder Ereignissen täuschend ähnlich sind. Betreiber – hier geht es nun um die Verwendung des AIS, nicht mehr seine Entwicklung – müssen in diesem Fall offenlegen, dass die Inhalte künstlich erzeugt oder manipuliert wurden (Art. 50 Abs. 4).

  Bei offensichtlich künstlerischen, kreativen, satirischen, fiktionalen oder analogen Werken muss der Hinweis auf die künstliche Herstellung oder Manipulation so erfolgen, dass die Darstellung oder den Genuss des Werks nicht beeinträchtigt wird.

- **Generative AIS**: Betreiber eines generativen AIS müssen offenlegen, dass die Inhalte künstlich erzeugt oder manipuliert wurden (Art. 50 Abs. 4). Das betrifft allerdings nur veröffentlichte Texte, wenn es um die Information der Öffentlichkeit über Angelegenheiten von öffentlichem Interesse geht, und nicht, soweit die generierten Texte menschlich überprüft oder redaktionell kontrolliert wurden und jemand die redaktionelle Verantwortung für die Veröffentlichung trägt. Eine Ausnahme gilt sodann für den Strafverfolgungsbereich.

- **Emotionserkennung**: Der Betreiber eines (nicht verbotenen → 27) Emotionserkennungssystems oder eines Systems zur biometrischen Kategorisierung müssen die betroffenen natürlichen Personen informieren (Art. 50 Abs. 3; wiederum mit einer Ausnahme für den Bereich der Strafverfolgung).
- **Entscheidungen**: Wenn der Betreiber eines HRAIS nach Anhang III (Use Cases → 28) das HRAIS verwendet, um eine Entscheidung zu fällen oder dabei zu unterstützen, die natürliche Personen betrifft, müssen diese entsprechend informiert werden (Art. 26 Abs. 11).
- **Menschliche Aufsicht**: Art. 26 enthält Vorgaben an Betreiber zur Wahrnehmung menschlicher Aufsicht, die ebenfalls einen Transparenzaspekt haben.

Die Pflichtinformation muss jeweils spätestens bei der der ersten Interaktion oder Aussetzung klar, eindeutig und barrierefrei bereitgestellt werden (Art. 50 Abs. 5).

Bei **GPAIM** werden ebenfalls Transparenzmassahmen vorgegeben, aber separat in Art. 53 (vgl. → 40 und 42). Weitere Anforderungen können nach anderen Vorgaben gelten, bspw. bei der Bearbeitung von Personendaten aus den Informations- und Transparenzpflichten des anwendbaren Datenschutzrechts.

38 Welche Anforderungen stellt der AI an “AI Literacy”?

“AI Literacy” bzw. “KI-Kompetenz” meint die für eine sachkundige und risikobewusste Verwendung eines AIS erforderlichen Fähigkeiten (Art. 3 Nr. 56). Art. 4 verlangt daher Massnahmen, um dem Personal und Hilfspersonen diese Kompetenz zu vermitteln (soweit sie mit einem AIS umgehen sollen). Dafür kommen vor allem Schulungen, Anleitungen und andere Informationen in Betracht.

Dieses “Upskilling” ist die einzige ausdrückliche Pflicht, die der AIA den Anbietern und Betreibern aller AIS auferlegt. Solche AIS können aber unter sektorielle Vorgaben fallen, und wenn sie an Konsumenten abgegeben werden, kann das allgemeine Produktesicherheitsrecht gelten. Ob das schweizerische PrHG auch für AIS gilt, die nicht in ein Produkt wie bspw. einen Roboter verbaut wurde, ist nicht abschliessend geklärt. Weitere Pflichten ergeben sich bei AIS in besonderen Konstellationen aus den Transparenzanforderungen (→ 37).

GPAI

39 Was ist ein AI-Modell mit allgemeinem Verwendungszweck (GPAIM)?

GPAIM werden im eigenen Kapitel V separat geregelt. Das ist der Gesetzgebungshistorie zuzuschreiben, bei der die Regelung von GPAI bis zuletzt strittig war (→ 3). Innerhalb der GPAI-Modelle wird eine besonders heikle Kategorie geregelt, die GPAI-Modelle "mit systemischen Risiken" (→ 41).

GPAIM sind "KI-Modelle" (kein definierter Begriff), die allgemein verwendbar sind, ein "breites Spektrum unterschiedlicher Aufgaben kompetent" erfüllen und die in nachgelagerte AIS integriert werden können (Art. 3 Nr. 63). Es geht dabei vor allem um Large Language Models (LLMs) wie bspw. ChatGPT oder Claude von Anthropic usw. Eine allgemeine Verwendbarkeit wird vermutet, wenn ein Modell mindestens eine Milliarde Parameter aufweist und mit einer grossen Datenmenge "unter umfassender Selbstüberwachung trainiert" wurde (ErwG 98 → 12). Kein GPAIM wäre dagegen ein Modell, bspw. ein LLM, das für einen engen Anwendungsbereich trainiert wurde.

Wesentlich ist dabei, dass ein GPAIM kein AIS ist. Es wird erst durch **Hinzufügung weiterer Komponenten** zum AIS und ggf. zum HRAIS (ErwG 97: "sollte klar bestimmt und vom Begriff der KI-Systeme abgegrenzt werden"; "obwohl KI-Modelle wesentliche Komponenten von KI-Systemen sind, stellen sie für sich genommen keine KI-Systeme dar"). Also: GPAI-Modell + weitere Komponente = AIS. Es braucht wenig für den Schritt vom GPAI-Modell zum (HR)AIS: Es reicht eine Nutzerschnittstelle (ErwG 63).

Möglich ist ferner auch, das ein GPAIM in ein anderes Modell verbaut und dieses dadurch zum GPAIM wird (ErwG 100). LLMs können auch weiter trainiert werden (z.B. durch Finetuning → 12). Wenn sich der Anwendungsbereich dadurch ausreichend verengt, ist es denkbar, dass das entsprechende Modell keine allgemeine Verwendbarkeit mehr hat.

Der Anbieter eines GPAIM – also diejenige Stelle, die das GPAIM entwickelt und in Verkehr bringt – wird also zum Anbieter des (HR)AIS, sobald er das GPAIM einem konkreten Einsatz zuführt und das resultierende AIS auf dem Markt bereitgestellt oder in Betrieb genommen wird. Dieser Logik folgend verlangt Art. 53 u.a., dass der GPAIM-Provider dem nachgelagerten AIS-Provider bestimmte **Informationen zur Verfügung stellen** muss (und zwar auch dann, wenn dieses kein HRAIS ist).

Ein LLM (→ 12) von OpenAI wäre ein Beispiel eines GPAI-Modells. ChatGPT verfügt dagegen über eine Nutzerschnittstelle und dürfte entsprechend ein AIS sein (auch wenn das nicht unstrittig ist). Verwendet ein Dritter ein Modell von OpenAI und baut damit seinen eigenen Chatbot, so ist dieser Dritte und nicht OpenAI der Anbieter des Chatbots als AIS. Das gilt selbstverständlich auch dann, wenn der entsprechende Dritte den Chatbot durch ein Finetuning weiter seinen eigenen Bedürfnissen anpasst.

Neben dem GPAIM definiert der AIA auch **GPAIS** (KI-Systeme mit allgemeinem Verwendungszweck; Art. 3 Nr. 66). GPAIS sind ein Unterfall von AIS und unterstehen den entsprechenden Regelungen. Der AIA erwähnt GPAIS deshalb nur am Rande (in Art. 3 Nr. 68, Art. 25 Abs. 1 lit. c, Art. 50 Abs. 2 und Art. 75 Abs. 2, und in einigen Erwägungsgründen).

40 Welche Pflichten haben Anbieter von GPAIM?

Die Pflichten des GPAIM-Anbieters (→ 20) werden wie erwähnt in einem **eigenen Kapitel** geregelt. Die Anforderungen an HRAIS-Anbieter – insbesondere Art. 16 AIA und die dort verwiesenen Bestimmungen – gelten für GPAIM-Anbieter nicht. GPAIM-Anbieter müssen aber:

- eine **technische Dokumentation** des GPAIM erstellen, einschliesslich des Trainings- und Testverfahrens und der Ergebnisse der Bewertung. Die Mindestinformationen legt Anhang XI fest. Auf Anfrage ist sie dem AI Office und den zuständigen nationalen Behörden zur Verfügung zu stellen (Art. 53 Abs. 1 lit. a). Davon gilt eine Ausnahme für FOSS (Art. 53 Abs. 2);
- weitere Informationen über das GPAIM dokumentieren (insbesondere nach Anhang XII) und diese den **Anbietern nachgelagerter AIS** zur Verfügung stellen (Art. 53 Abs. 1 lit. b). Davon gilt ebenfalls die Ausnahme für FOSS (Art. 53 Abs. 2);
- über eine Strategie zur **Einhaltung des EU-Urheberrechts** verfügen. Dazu gehört auch eine Angabe dazu, wie im Fall der Text and data Mining-Ausnahme (→ 59) ein Nutzungsvorbehalt i.S.v. Art. 4 Abs. 3 der Urheberrechts-Richtlinie (https://dtn.re/c6zFb9) eingehalten wird (Art. 53 Abs. 1 lit. c). Dabei ist zu beachten, dass diese Anforderung nach ErwG 106 wohl auch für aussereuropäische GPAIM-Anbieter gilt, die ein GPAIM in der EU in Verkehr bringen;
- eine Zusammenfassung der **Trainingsdaten** veröffentlichen (dafür soll das AI Office eine Vorlage ausarbeiten), unter Vorbehalt von Geschäftsgeheimnissen (ErwG 107).

Sie müssen ferner ggf. einen **Bevollmächtigten** bestellen (Art. 54 AIA → 26). Wie anderswo kann die Kommission die Anforderungen weiter konkretisieren (→ 51).

41 Wie regelt der AIA GPAIM mit systemischen Risiken?

Systemische Risiken sind Risiken, die sich aufgrund der "Reichweite" des GPAIM oder aufgrund möglicher negativer Folgen "für die öffentliche Gesundheit, die Sicherheit, die öffentliche Sicherheit, die Grundrechte oder die Gesellschaft insgesamt" erhebliche Auswirkungen haben und sich über die gesamte Wertschöpfungskette hinweg verbreiten können (Art. 3 Nr. 65).

Ob dies auf ein GPAIM zutrifft, entscheidet sich allerdings nicht nach der Legaldefinition, sondern nach den Kriterien von Art. 51 Abs. 1. Ein systemisches Risiko liegt danach in zwei Fällen vor:

- Wenn das GPAIM über "**Fähigkeiten mit hohem Wirkungsgrad**" verfügt, was durch geeigneter Methoden wie bspw. Benchmarks zu bewerten ist (Art. 51 Abs. 1 lit. a), aber jedenfalls dann vorliegt, wenn das "die kumulierte Menge der für sein Training verwendeten Berechnungen" mehr als 1025 Gleitkommaoperationen beträgt. Gleitkommaoperationen wiederum werden in Art. 3 Nr. 67 als mathematische Grösse definiert. Dieser Schwellenwert dürfte in Zukunft angepasst werden (ErwG 111).
- wenn die **EU-Kommission entscheidet**, dass ein systemisches Risiko besteht, wobei Anhang XIII die entsprechenden Kriterien liefert (lit. b und Art. 52 Abs. 4-5). Es geht um die Leistungsfähigkeit des Modells, ausgedrückt u.a. durch die Anzahl Parameter oder den Umfang der Trainingsdaten, aber auch die Grösse des Markts des Modells.

Der Anbieter muss das GPAIM mit systemischen Risiken zunächst der **Kommission melden** (→ 51), so bald wie möglich, wenn das GPAIM die Schwelle der systemischen Risiken erreicht hat, spätestens aber nach zwei Wochen (Art. 52 Abs.

1). Er kann sodann nachzuweisen versuchen, dass sein GPAIM ausnahmsweise dennoch keine systemischen Risiken mit sich bringt, wenn die initiale Qualifikation auf dem materiellen Kriterium von Art. 51 Abs. 1 lit. a beruht. Er muss der Kommission dazu entsprechende Argumente vortragen. Überzeugt er die Kommission nicht, wird das GPAIM auf der Liste der systemischen Risiken eingetragen (Art. 51 Abs. 3). – Hat die Kommission das GPAIM von Amts wegen als systemisch riskant eingestuft, kann der Anbieter jederzeit Wiedererwägung verlangen (Art. 51 Abs. 5).

42 Welche Pflichten haben Anbieter von GPAIM mit systemischen Risiken?

Anbieter vom GPAIM **mit systemischen Risiken** haben zusätzliche Pflichten, d.h. zusätzlich zu den Pflichten der Anbieter von weniger heiklem GPAIM. Sie müssen (Art. 55):

- das GPAIM standardisiert bewerten;
- systemische Risiken auf Ebene EU bewerten und reduzieren;
- Informationen über schwerwiegende Vorfälle und mögliche Abhilfemassnahmen dokumentieren und ggf. das AI Office und die zuständigen nationalen Behörden informieren; und
- ausreichende Cybersicherheit gewährleisten.

AIS im Betrieb

43 Wie ist die Marktüberwachung geregelt?

Die Marktüberwachung ist ein zentrales Element des AIA – sie soll sowohl die Compliance von AIS im Interesse der betroffenen Personen als auch ein Level Playing Field sicherstellen.

Anbieter müssen deshalb ein **System zur Marktbeobachtung** nach dem Inverkehrbringen von HRAIS einrichten und dokumentieren (Art. 71 Abs. 1). Dazu gehört die Erhebung, Dokumentation und Auswertung von Daten über die Leistung der HRAIS (die u.U. über die Betreiber beschafft werden) während des gesamten Lifecycle des HRAIS.

Dieses System umfasst insbesondere einen **Plan für die Beobachtung** der HRAIS nach ihrem Inverkehrbringen. Dieser Plan wiederum ist Teil der technischen Dokumentation nach Anhang 4 (→ 35 Nr. 6); die Kommission soll noch festlegen, wie ein solcher Plan aussehen soll (Art. 72 Abs. 3). Soweit ein HRAIS unter Anhang I Abschnitt A fällt (bspw. Medizinprodukte), können Anbieter die Anforderungen des AIA auch in den schon vorhandenen Systemen und Plänen integrieren (Art. 72 Abs. 4).

Ebenfalls zur Marktüberwachung gehören die Pflicht, bei **Non-Compliance** zu reagieren (→ 44), bestimmte **Vorfälle** zu melden (→ 45), und die entsprechenden Befugnisse der Behörden.

Allgemeiner stellen AIS grundsätzlich auch Produkte im Sinne der **Marktüberwachungsverordnung** dar (Art. 74 Abs. 1; https://dtn.re/JgakBQ). Die Marktüberwachungsbehörden (→ 55) können daher aktiv werden, wann immer ein AIS – es muss kein HRAIS sein – wahrscheinlich die Gesundheit oder Sicherheit der Nutzer gefährdet anwendbaren Harmonisierungsrechtsvorschriften nicht entspricht (Art. 16 Abs. 1 der Marktüberwachungsverordnung).

44 Was gilt, wenn ein HRAIS nicht (mehr) compliant ist?

Nicht erst bei schwerwiegenden Vorfällen muss reagiert werden (→ 45), sondern selbstverständlich auch immer dann, wenn ein HRAIS den entsprechenden Anforderungen nicht mehr entspricht. Dabei nimmt der AIA nicht nur den Anbieter in die Pflicht, sondern auch weitere Akteure.

Haben Anbieter Grund zur Annahme, dass ein HRAIS nicht mehr dem AIA entspricht, zu irgendeinem Zeitpunkt nach dem Inverkehrbringen bzw. der Inbetriebnahme, müssen sie den Mangel sofort beheben oder das HRAIS ggf. vom Markt **zurückzunehmen, deaktivieren oder zurückrufen** (Art. 20 Abs. 1). “Rücknahme” meint dabei, die Bereitstellung eines bereits in der Lieferkette befindlichen HRAIS verhindert wird (Art. 3 Nr. 17), und “Rückruf” heisst, dass HRAIS zurückgegeben oder wenigstens ausser Betrieb gesetzt oder abgeschaltet werden (Art. 3 Nr. 16).

Anbieter müssen auch den **nachgelagerten Markt** entsprechend informieren, d.h. die Händler, die Betreiber, den Bevollmächtigten und die Einführer (Art. 20 Abs. 1). Sofern das HRAIS zudem ein Risiko nach Art. 79 Abs. 1 AIA birgt, gelten die entsprechenden Pflichten (→ 45).

Die nachgelagerten Akteure werden im Fall der Non-Compliance ebenfalls einbezogen. Einführer dürfen das HRAIS in diesem Fall erst in Verkehr bringen, wenn die Konformität wiederhergestellt wurde (Art. 23 Abs. 2), und dasselbe gilt für Händler mit Bezug auf die Bereitstellung auf dem Markt (Art. 24 Abs. 23).

Auch **Bevollmächtige** haben Aufgaben: Wenn sie Grund zu der Annahme haben, dass der Anbieter gegen den AIA verstösst, müssen sie ihr Mandat beenden und die zuständige Marktüberwachungsbehörde und ggf. die notifizierte Stelle darüber und über die Gründe informieren (Art. 22 Abs. 4).

45 Wie ist mit Incidents und mit besonderen Risiken umzugehen?

Als Teil der Marktüberwachung (→ 43) müssen bestimmte Vorfälle dokumentiert und gemeldet werden. Diese Pflicht trifft die Anbieter von HRAIS, und wird durch "**schwerwiegende Vorfälle**" ausgelöst. Das sind Fehlfunktionen, aber auch generell Vorfälle, die direkt oder indirekt zum Tod oder zu einer schweren gesundheitlichen Schädigung führen, zu einer "schweren und unumkehrbaren" Störung der Verwaltung oder des Betriebs kritischer Infrastrukturen, zu einer Verletzung von Grundrechten oder zu schweren Sach- oder Umweltschäden (Art. 3 Nr. 49).

Tritt ein solcher Vorfall ein, muss der Anbieter den Vorfall den zuständigen **Marktüberwachungsbehörden** melden (→ 54), wobei Sonderregeln für gewisse HRAIS gelten. Die Meldung muss sofort bei Feststellung durch den Anbieter erfolgen, spätestens aber **15 Tage** nach Kenntnis durch den Anbieter oder auch durch den Betreiber (Art. 73 Abs. 2).

Soweit ein Vorfall breite Auswirkungen hat ("widespread infringement") oder eine kritische Infrastruktur betrifft, verkürzt sich die Meldefrist auf **zwei Tage** (Art. 73 Abs. 3 AIA), und im Fall des Todes auf **zehn Tage** (Abs. 4). Wie im Datenschutzrecht oder bei Meldungen gegenüber der FINMA kann mit einer Erstmeldung und einer Folgemeldung gearbeitet werden.

Nach der Meldung informieren die Marktüberwachungsbehörden die zuständigen nationalen Behörden. Sollte es erforderlich sein, müssen sie zudem innerhalb von sieben Tagen anordnen, dass das HRAIS zurückgerufen bzw. vom Markt genommen oder dass die Bereitstellung auf dem Markt untersagt wird (Art. 73 Abs. 8 i.V.m. Art. 19 der Marktüberwachungsverordnung (https://dtn.re/ElQE2G).

Der Anbieter muss den Vorfall ferner untersuchen und die Risiken einschätzen und nach Möglichkeit **mitigieren** (Art. 73 Abs. 6 AIA), in Zusammenarbeit mit den zuständigen Behörden.

Neben den Anbietern haben auch **Betreiber** Pflichten im Fall eines schwerwiegenden Vorfalls. Sie müssen solche Vorfälle dem **Anbieter mitteilen** (Art. 26 Abs. 5 und Art. 72). Bei besonders heiklen HRAIS oder beim Einsatz in kritischen Infrastrukturen sind in der Praxis wohl vertragliche Regelungen dieser Meldepflicht zu erwarten, auch wenn sie sich bereits aus dem AIA ergibt.

Von den schwerwiegenden Vorfällen ist der Fall zu unterscheiden, dass ein HRAIS zu besonderen Risiken führt, d.h. **atypisch hohen Risiken** für die Gesundheit oder Sicherheit oder Grundrechte (Art. 79 Abs. 1). In diesem Fall haben verschiedene Rollen entsprechende Pflichten. Hat eine Marktüberwachungsbehörde Grund zur Annahme, dass solche Risiken vorliegen, prüft sie das betreffende AIS und – sollte sich die Annahme bestätigen – informiert die zuständigen nationalen Behörden. Auch **Betreiber** haben bei einem solchen System besondere Pflichten, falls es sich um ein HRAIS handelt.

46 Welche Rechte haben Betroffene und andere Stellen?

Alle Personen (natürliche und juristische) haben das Recht, sich bei der zuständigen Marktüberwachungsbehörde (→ 55) zu beschweren, wenn sie Grund zur Annahme haben, dass eine Bestimmung des AIA verletzt wurde (Art. 85 Abs. 1). Dazu muss eine Person nicht besonders betroffen sein – auch Konkurrentenbeschwerden sind möglich.

Von einer erheblichen **Entscheidung** betroffene Personen haben weiter das Recht, vom Betreiber eine Erläuterung zur Rolle des AIS bei der Entscheidung und zu den Kernelementen der Entscheidung zu verlangen (→ 35 Nr. 13).

Betroffene haben weiter das Recht, eine Beschwerde ans AI Office zu richten (Art. 89 Abs. 2). Das gilt auch für Anbieter, die ein GPAIM in ein eigenes AIS verbaut haben.

Dazu kommen Rechte nach anderen Rechtsgrundlagen, insbesondere auch nach dem anwendbaren **Datenschutzrecht** (→ 58) und ggf. nach vertraglichen Regelungen. Auch Schadenersatzansprüche kommen bei gegebenen Voraussetzungen in Frage.

Sonderfragen

47 Werden KMU bei der Anwendung des AIA entlastet?

Für KMU wird die Umsetzung der Anforderungen des AIA herausfordernd sein, zumindest soweit sie als Anbieter tätig sind. Wer ein GPAIM einkauft und als HRAIS in Verkehr bringt, wird dadurch zum Anbieter des HRAIS – es dürfte also eine recht grosse Zahl von KMUs geben, die auf Basis eines LLMs einen bestimmten Use Case abdecken und für diesen Anbieter sind.

Grundsätzlich gelten die Bestimmungen des AIA tel quel auch für KMU. Der AIA enthält aber einige Bestimmungen, die KMU bzw. SME in der englischen Version unterstützen sollen:

- Art. 62 verpflichtet die Mitgliedstaaten zu Förderungsmassnahmen, indem KMU prioritärer Zugang zu KI-Reallaboren gewähren werden soll, Sensibilisierungs- und Schulungsmassnahmen für KMU durchzuführen sind, Fragen zum AIA und zu KI-Reallaboren zur werden können und KMU bei der Entwicklung von Normen (→ 15) einbezogen werden sollen.
- KMU sollen beim Advisory Forum mitwirken (Art. 67 Abs. 2).
- Bei Verhaltenskodizes sind die Interessen der KMU zu berücksichtigen (Art. 95 Abs. 4).
- Bei den Bussen gilt ein etwas niedrigerer Ansatz (Art. 99 Abs. 6).

Für Kleinstunternehmen im Sinne der Kommissionsempfehlung K(2003)1422 (https://dtn.re/U7vlKH) sieht Art. 63 Abs. 1 zudem eine Erleichterung beim QMS (→ 35) vor.

48 Was sind KI-Reallabore und Tests unter Realbedingungen?

Der AIA schreibt sich die Innovationsförderung in diversen Erwägungsgründen auf die Fahnen, und sein grösster Beitrag zur Innovationsförderung besteht wohl darin, dass er kein Verbotsgesetz ist (mit den wenigen Ausnahmen, Q0). Das Kapitel VI (Art. 57 ff.) ist sodann ausdrücklich der Innovationsförderung gewidmet.

Dazu dienen im Wesentlichen zwei Elemente. Das erste Element sind die "**KI-Reallabore**" (der entsprechende englische Begriff ist "AI regulatory sandbox"):

- Dabei handelt es sich um eine Erleichterung der Entwicklung, des Trainings, des Testens und der Validierung von AIS vor dem Inverkehrbringen oder der Inbetriebnahme nach einem Plan, der zwischen den Anbietern und der zuständigen Behörde zu vereinbaren ist (Art. 57 Abs. 5), und ggf. unter Beizug der Datenschutzbehörden (Abs. 10).
- Art. 59 enthält sodann eine beschränkte Rechtsgrundlage für die Bearbeitung von **Personendaten** im Rahmen eines Reallabors: Personendaten dürfen für die Entwicklung, das Training und für Tests im Reallabor verarbeitet werden, allerdings nur, wenn bestimmte Bedingungen erfüllt sind und nur bei der Entwicklung eines AIS zur Wahrung bestimmter öffentlichen Interessen. Diese Rechtsgrundlage tritt neben die analoge Rechtsgrundlage für Testzwecke nach Art. 10 (→ 36).
- Anbieter können anschliessend einen Nachweis über die im Reallabor durchgeführten Tätigkeiten und einen Abschlussbericht erhalten, was das **Konformitätsbewertungsverfahren** oder die Marktüberwachung erleichtern soll (Abs. 7). Die Einhaltung des Plans bietet zudem einen Safe Harbor gegen Bussen im Fall einer mit dem Plan zusammenhängen Verlet-

zung des AIA, ggf. aber auch anderer Vorgaben insbesondere auch des Datenschutzrechts (Abs. 12).

– Jeder Mitgliedstaat muss bis am 2. August 2026 mindestens ein solches Labor einrichten (Art. 57 Abs. 1). Die Kommission soll vorher aber noch detailliertere Regelungen erlassen (Art. 58 AIA).

Das zweite Element sind Tests von Anhang III-HRAIS unter **Realbedingungen**:

– HRAIS nach Anhang III (d.h. die use case-bezogenen HRAIS; Q28) können unter bestimmten Voraussetzungen ausserhalb eines KI-Reallabors unter Realbedingungen durchgeführt werden (Art. 60). Vorausgesetzt ist, dass der Test kontrollierbar ist, d.h. dass der Test wirksam überwacht wird und Vorhersagen, Empfehlungen oder Entscheidungen des AIS rückgängig gemacht oder ausser Acht gelassen werden können (Art. 60 Abs. 4 lit. j-k). Schwerwiegende Vorfälle sind nach Art. 73 zu melden, d.h. die entsprechende Meldepflicht (→ 45) wird auf den Zeitpunkt vor dem Inverkehrbringen bzw. der Inbetriebnahme vorbezogen (Art. 60 Abs. 7).

– Tests müssen auf einem Plan beruhen, der von der zuständigen Marktüberwachungsbehörde zu genehmigen ist (Art. 60 Abs. 4 lit. a-b).

– Soweit der Plan die Teilnahme von Testteilnehmern erfordert, müssen diese in die Teilnahme grundsätzlich einwilligen (Art. 61 Abs. 4 lit. j und Abs. 5).

Sanktionen und Governance

49 Was gilt bei Verletzungen des AIA?

Kapitel XII betrifft die Sanktionen bei Verletzungen des AIA. Der AIA selbst enthält – anders als die DSGVO – keine eigenen Bussgeldtatbestände, sondern verpflichtet die Mitgliedstaaten in Art. 99 zur Einführung Bestimmungen über Bussen, aber auch andere Durchsetzungsmassnahmen. Bussen können sich gegen alle Akteure richten, also alle in der Wertschöpfung beteiligten Stellen.

Die Bussen können je nach Art des Verstosses bis zu EUR 35 Mio. EUR bzw. 7% des Umsatzes erreichen:

- Bei einer Verletzung der **verbotenen Praktiken** (→ 27) gilt der obere Bussenbetrag von bis zu EUR 35 Mio. oder bei 7% des weltweiten Jahresumsatzes (Art. 99 Abs. 3). Wie bei der DSGVO dürfte dafür der Konzernumsatz massgebend sein.
- Bei bestimmten **anderen Verletzungen** liegt der obere Bussenrahmen bei EUR 15 Mio. oder bei 3% des Jahresumsatzes (Art. 99 Abs. 4). Diese Bussen können sich gegen Akteure, aber auch notifizierte Stellen richten. Das betrifft Verletzungen von Art. 16 (Anbieter), Art. 22 (Bevollmächtigte), Art. 23 (Einführer), Art. 24 (Händler), Art. 26 (Betreiber) und Art. 31, 33 Abs. 1, 3 und 4 und Art. 34 (notifizierte Stellen) sowie Art. 50 (Transparenz; Anbieter und Betreiber).
- Im Fall von **falschen Antworten** an notifizierte Stellen oder die zuständigen nationalen Behörden liegt der Bussenrahmen bei EUR 7.5 Mio. oder bei 1% des Jahresumsatzeses (Art. 99 Abs. 5).

Massgebend ist jeweils der höhere Betrag, ausser bei KMU (hier der niedrigere; Art. 99 Abs. 6; → 47). Im konkreten Fall hat haben das Gericht oder die Verwaltungsbehörde (Art. 99 Abs. 9) bei der Bussenbemessung die Kriterien von Art. 99 Abs. 7 zu berücksichtigen, u.a. die Schwere des Verschuldens.

Bei **Anbietern von GPAIM** enthält Art. 101 eine Sonderregelung. Alle Verletzungen des AIA können hier mit Busse geahndet werden (Art. 101 Abs. 1 lit. a); dennoch nennt Art. 101 Abs. 1 eigens bestimmte Verletzungen. Der Bussenrahmen liegt hier bei EUR 15 Mio. oder zu 3% des Jahresumsatzes.

Von Verletzungen zu unterscheiden ist natürlich der Eintritt eines schwerwiegenden Vorfalls (→ 45).

50 Welche Behörden spielen beim AIA eine Rolle?

Der AIA regelt die Rolle mehrerer Behörden vorwiegend im eigenen Kapitel “Governance” (Kapitel VII, Art. 64 ff.). Diverse Behörden und Einrichtungen sind mit unterschiedlichen und teils überschneidenden Aufgaben betraut. Dabei gibt es sowohl eine horizontale Arbeitsteilung (innerhalb der EU) als auch eine vertikale (zwischen der EU und den Mitgliedstaaten).

Erstere regelt Abschnitt 1 des Kapitels VII (Governance). Die **Kommission** nimmt bei den Gremien der EU die führende Rolle ein, und ihr obliegt grundsätzlich die Durchsetzung des AIA. Sie hat weitreichende Befugnisse, kann konkretisierende Bestimmung erlassen und ist zuständig zur Entgegennahme von Mitteilungen der Akteure und anderer Behörden (→ 51).

Das **AI Office** (“Büro für Künstliche Intelligenz”) ist ein Teil der Kommission und zuständig für die Marktaufsicht von GPAIM und von AIS, die auf GPAIM desselben Providers basieren (Art. 88 und 75; Q52).

Das European AI Board (**EAIB**) soll die Kommission (und die Mitgliedsstaaten) dabei beraten und unterstützen (→ 53).

Die nationalen **Marktüberwachungsbehörden** sind zuständig für die Überwachung der Einhaltung des AIA (→ 54).

Die **notifizierenden nationalen Behörden** sind zuständig für die Bewertung, Benennung, Notifizierung und Überwachung von AI-Konformitätsbewertungsstellen (Art. 28).

Konformitätsbewertungsstellen sind wiederum Stellen, die die Konformität von AIS in Übereinstimmung mit dem AIA prüfen und bewerten (Art. 3 Ziff. 21 → 15).

Aufgrund der umfassenden Kooperationspflichten der Akteure und den weitreichenden Informationsbeschaffungsmöglichkeiten der Behörden werden die Kommission, die Marktüberwachungsbehörden und die notifizierten Stellen und alle anderen an der Anwendung des AIAI beteiligten Stellen einer **Vertraulichkeitspflicht** unterworfen (Art. 78).

51 Welche Aufgaben hat die EU-Kommission im Rahmen des AIA?

Die Hauptrolle auf Ebene der EU liegen bei der Kommission und beim AI Office als Teil der Kommission (→ 52).

Die Kommission, die nach Art. 17 Abs. 1 des Vertrags über die Europäische Union die Einhaltung des EU-Rechts überwacht, hat dabei eine zentrale Rolle. Ihre Befugnisse lassen sich wie folgt einteilen (nicht ganz vollständig – einige weitere, untergeordnete Aufgaben der Kommission werden nicht aufgeführt):

Konkretisierende Legiferierung: Art. 97 AIA überträgt der Kommission auf Basis von Art. 290 des EU-Vertrags (https://dtn.re/9MhpKX) das Recht, verbindliche "delegierte Rechtsakte" zu erlassen. Der EU-Vertrag unterscheidet zwischen delegierten Rechtsakten und Durchführungsrechtsakten. **Delegierte Rechtsakte** sind Rechtsakte zur Ergänzung oder Änderung des Basis-Rechtsakts (des AIA), die die Kommission dem Rat und dem Parlament zur Zustimmung oder Ablehnung vorlegt. **Durchführungsrechtsakte** sind blosse Umsetzungsbestimmungen wie technische Bestimmungen, Ausnahmen usw., die dem Parlament und Rat nicht vorgelegt werden.

Die Befugnis, **delegierte Rechtsakte** zu erlassen, beruht auf Art. 97 AIA und soll es ermöglichen, die im Bereich der AI besonders rasche technische Entwicklung abzubilden. Das betrifft abschliessend die folgenden Punkte:

- die Kriterien, wann ein AIS zum HRAIS wird (→ 28), und analog auch des Anhangs III (Use Cases; Art. 7 Abs. 1 und 3);
- den Anhang IV zum Mindestinhalt der technischen Dokumentation (Art. 11 Abs. 3);
- die Anhänge VI und VII und von Art. 43 Abs. 1 und 2 über das Konformitätsbewertungsverfahren und von Anhang V zum Inhalt der EU-Konformitätserklärung;
- die Kriterien für die Einstufung eines GPAIM als systemisch riskant nach Art. 51 Abs. 1 und 2 und Anhang XIII;
- die Anhänge XI und XII zum Inhalt der technischen Dokumentation und der Transparenzanforderungen bei der Downstream-Verwendung von GPAIM (Art. 53).

Daneben kann die Kommission in den folgenden Bereichen **Durchführungsrechtsakte** erlassen. Sie hat sich dabei i.d.R. nach der Durchführungsbefugnis-Verordnung (https://dtn.re/B9uV04) zu richten (Art. 98 Abs. 2):

- Eingreifen, wenn ein Mitgliedstaat die Anforderungen für **notifizierte Stellen** nicht erfüllt (Art. 37 Abs. 4);
- Genehmigung von **Praxisleitfäden** im Zusammenhang mit GPAIM gemäss Art. 56, generell und insbesondere auch zur Konkretisierung der Transparenzanforderungen bei AI-generierten oder manipulierten Inhalten (Art. 50 Abs. 7), der Pflichten der Anbieter von GPAIM nach Art. 53 und von systemisch riskanten GPAIM nach Art. 55 (Art. 56 Abs. 6);

- Erlass **gemeinsamer Spezifikationen**, wenn einschlägige Normen fehlen (Art. 41 AIA), und gemeinsamer Vorschriften im Bereich der GPAIM, wenn bis zum 2. August 2025 kein Verhaltenskodex vorliegt (Art. 56 Abs. 9)
- Konkretisierende Regelungen für **KI-Reallabore** (Art. 58 Abs. 1 und 2) und Tests von HRAIS unter Realbedingungen (Art. 60);
- Bestimmungen zur Einrichtung eines **wissenschaftlichen Gremiums** unabhängiger Sachverständiger (Art. 68 Abs. 1 und 5 und Art. 69 Abs. 2);
- Konkretisierungen für den **Marktbeobachtungsplan** der Anbieter von HRAIS (Art. 72 Abs. 3);
- Konkretisierungen des **Sanktionsverfahrens** (Art. 101 Abs. 6).

Die Kommission kann sodann durch den **Erlass von Leitlinien und Normung** zur Vereinheitlichung der Praxis beitragen**:**

- Die Kommission soll die Fäden bei der Anwendung des AIA generell in der Hand halten. Sie erteilt bspw. **Normungsaufträge** gemäss Art. 10 der Normungsverordnung, (https://dtn.re/BRL10Q), also Aufträge zur Ausarbeitung derjeniger Normen, deren Einhaltung eine Konformitätsvermutung begründet (Art. 40 Abs. 1 und 2), und kann – wenn einschlägige Normen fehlen – entsprechende "gemeinsame Spezifikationen" erlassen (Art. 41 AIA).
- Nach Art. 96 AIA kann sie ferner allgemein **Leitlinien** für die praktische Umsetzung des AIA erlassen. Art. 96 enthält zwar eine Liste zu konkretisierender Punkte – besonders die Definition des AIS, die Anwendung der Art. 8 ff. mit den grundlegenden Anforderungen, die Einstufung als HRAIS (Art. 6 Abs. 5), die verbotenen Praktiken und die Transparenz nach Art. 50 AIA – , aber sie ist nicht abschliessend.
- Die Kommission **genehmigt** ferner **Praxisleitfäden** nach Art. 56 AIA, d.h. eine Konkretisierung der Pflichten der Anbieter von GPAIM.
- Sie stellt auch **Vorlagen und Formulare** zur Verfügung, die in der Praxis von erheblicher Bedeutung sein dürften. Sie soll ein vereinfachtes Formular für die technische Dokumentation bei HRAIS von KMU zur Verfügung stellen (Art. 11; Anhang IV).

Die Kommission nimmt weiter **Meldungen und Berichte** entgegen:

- biometrische Echtzeit-Fernidentifizierungen zu Strafverfolgungszwecken: Mitteilung der Mitgliedstaaten über entsprechende Rechtsgrundlagen (Art. 5 Abs. 5) und Jahresberichte der nationalen Marktüberwachungs- und Datenschutzbehörden (Art. 5 Abs. 6);
- Konformitätsbewertung: Mitteilung der notifizierenden Behörden über Konformitätsbewertungsstellen (Art. 30 Abs. 2 f. und Art. 36 Abs. 1, 4 und 7); Mitteilung der Marktüberwachungsbehörden über Ausnahmebewilligungen für HRAIS nach Art. 46 Abs. 1 (Art. 46 Abs. 3; wobei die Kommission einschreiten kann);
- GPAIM: Meldung der Anbieter von GPAIM mit systemischen Risiken (Art. 52 Abs. 1);
- Mitteilung der notifizierenden Behörden und Marktüberwachungsbehörden durch die Mitgliedstaaten (Art. 70 Abs. 2 und 6);
- Mitteilung der nationalen Behörden über schwerwiegende Vorfälle (Art. 73 Abs. 11) gemäss der Marktüberwachungsverordnung; (https://dtn.re/ubfelK);
- jährliche Mitteilung der Marktüberwachungsbehörden über die Informationen aus der Marktüberwachung und über die Anwendung verbotener Praktiken (Art. 74 Abs. 2);
- Mitteilung der Mitgliedstaaten über die nationale Behörden oder öffentlichen Stellen für die Aufsicht über den Schutz der Grundrechte (Art. 77 Abs. 1 und 2);
- Informationen der Mitgliedstaaten im Zusammenhang mit riskanten AIS i.S.v. Art. 79 Abs. 1(Art. 79 Abs. 3 ff.);
- Informationen der Mitgliedstaaten im Zusammenhang mit riskanten AIS, die der Anbieter als nicht hochriskant eingestuft hat (Art. 80 Abs. 3), und mit konformen HRAIS, die dennoch ein besonderes Risiko mit sich bringen (Art. 82 Abs. 1 und 3);

- Mitteilungen der Mitgliedstaaten über ihre Sanktions- und sonstigen Durchsetzungsbestimmungen und über ihre Bussenpraxis (Art. 99 Abs. 2 und 11); Mitteilung des EDPS über seine Bussenpraxis (Art. 100 Abs. 7).

Die Kommission hat weiter **Eingriffs- und Entscheidungsbefugnisse**:

- Sanktionierung von Anbietern von GPAIM (Art. 101 Abs. 1);
- Einwände gegen Ausnahmebewilligungen für HRAIS nach Art. 46 Abs. 1 (Art. 46 Abs. 4 und 5);
- Einstufung eines GPAIM als systemisch riskant (Art. 52 Abs. 2-5);
- Bewertung der Verfahren, die Anbieter von GPAIM bzw. von systemisch riskanten GPAIM anwenden können, um den Nachweis ihrer jeweiligen Pflichten nach Art. 53 bzw. 55 zu führen (soweit keine harmonisierte Normen bestehen; Art. 53 Abs. 4 und 55 Abs. 2);
- Eingreifen, wenn ein AIS mit besonderen Risiken i.S.v. Art. 79 Abs. 1 nicht konform ist oder ein konformes HRAIS dennoch besonders riskant ist und die Kommission mit den Massnahmen der zuständigen Marktüberwachungsbehörde nicht einverstanden ist (Art. 81 und 82).

Schliesslich informiert die Kommission durch **Veröffentlichungen und Bekanntmachungen**:

- Verzeichnis der notifizierenden Stellen (Art. 35 Abs. 2);
- Liste von systemisch riskanten GPAIM (Art. 52 Abs. 6);
- Liste der zentralen Anlaufstellen der Mitgliedstaaten (Art. 70 Abs. 2);
- Datenbank der HRAIS nach Anhang III (Art. 71);
- Berichterstattung an das Parlament und den Rat (Art. 112).

Und zuletzt hat die Kommission **Durchsetzungsbefugnisse bei GPAIM**:

- GPAIM werden im Kapitel V besonders geregelt. Die Kommission ist beauftragt, die Vorschriften dieses Kapitels durchzusetzen; dies regelt der eigene Abschnitt 5 des Kapitels IX (Beobachtung nach den Inverkehrbringen; Informationsaustausch und Marktüberwachung). Sie ist von den Marktüberwachungsbehörden entsprechend auf dem Laufenden zu halten (Art. 73 Abs. 11, Art. 74 Abs. 2, Art. 77 Abs. 2, Art. 79 Abs. 3 ff., Art. 80 Abs. 3).
- Die Kommission kann eingreifen, wenn sie mit Massnahmen der Mitgliedstaaten, wenn AIS oder HRAIS mit besonderen Risiken nicht einverstanden ist (Art. 81 und Art. 82 Abs. 4 f.).
- Sie ist ferner generell zuständig, das Kapitel V durchzusetzen (Art. 88 Abs. 1). Zu diesem Zweck kann sie Informationen von den Anbietern von GPAIM anfordern (Art. 91 Abs. 1 und 3 und Art. 92 Abs. 3), Sachverständige für die Bewertung von GPAIM einsetzen (Art. 92 Abs. 2) und Anbieter von GPAIM auffordern, ihre Pflichten einzuhalten, Risikominderungsmassnahmen zu treffen und ein GPAIM vom Markt zu nehmen (Art. 93 Abs. 1).

52 Welche Rolle hat das AI Office?

Das AI Office (https://dtn.re/AkFqHT) wird in Art. 3 Nr. 47 als “Büro für Künstliche Intelligenz” legaldefiniert, und zwar nicht als Behörde, sondern als die Aufgabe der Kommission, zur Umsetzung, Beobachtung und Überwachung von AIS und GPAIM beizutragen. Ursprünglich war dafür eine eigene Behörde vorgesehen. Die Kommission hat dieses Büro mit Beschluss vom 24. Januar 2024 (https://dtn.re/cvmxvL) eingesetzt, allerdings mit leicht abweichender Bezeichnung, nämlich als “**Europäisches Amt für künstliche Intelligenz**”; beides ist das AI Office (das AI-Denglisch setzt sich durch). Es ist Teil der GD (Generaldirektion) Kommunikationsnetze, Inhalte und Technologien in der Kommission. Es hat mehr als 140 Mitarbeitende und ist in fünf Abteilungen aufgeteilt, “Excellence in AI and Robotics

Unit", "Regulation and Compliance Unit", "AI Safety Unit", "AI Innovation and Policy Coordination Unit" und "AI for Societal Good Unit".

Die Aufgaben des Amts ergeben sich aus Art. 3 Nr. 47, Art. 64, weiteren Bestimmungen des AIA und aus dem erwähnten Beschluss, der diese und weitere Aufgaben und die Befugnisse des Amts auflistet. Es geht vor allem um folgende:

- **koordinative Aufgaben** (bspw. die Zusammenarbeit mit Stakeholdern, den weiteren Abteilungen der Kommission, den weiteren Gremien der EU und mit den Mitgliedstaaten und ihren Behörden);
- **Fachbeiträge** (bspw. der Beobachtung der wirtschaftlichen und technischen Entwicklungen, der Ausarbeitung von Leitfäden und Musterbedingungen [Art. 25, 27 Abs. 5, 50 Abs. 7, 53 Abs. 1 lit. d, 56 und 62 Abs. 2] und der Vorbereitung von Beschlüssen der Kommission (Art. 56);
- die **Marktaufsicht** über GPAIM und über AIS, die ein Anbieter auf Basis eines eigenen GPAIM baut (Art. 88 und Art. 75 und Art. 3 des genannten Kommissionsbeschlusses). Es überprüft hier die Einhaltung des AIA durch die entsprechenden Akteure und dient auch als Anlaufstelle für die Meldung schwerwiegender Vorfälle (→ 45).

Daneben beaufsichtigt das Amt auch den AI Pact (https://dtn.re/WJfwxl).

53 Welche Rolle hat das EAIB?

Das "Europäische Gremium für Künstliche Intelligenz" ("KI-Gremium"; auch "**EAIB**", für "European AI Board", https://dtn.re/hF9LMM) setzt sich aus Vertretern der Mitgliedsstaaten zusammen, mit dem EDPS und Vertretern der weiteren EWR-Staaten als Beobachter und unter Beisitz eines Vertreters des AI Office (Art. 65 Abs. 1 und 2); im Wesentlichen analog zum EDPB bei der DSGVO. Am 10. September 2024 hat seine erste Sitzung stattgefunden (https://dtn.re/QQhGJ7).

Es soll die Kommission und die Mitgliedsstaaten beraten und unterstützen, um die einheitliche und wirksame Anwendung des AIA zu erleichtern (Art. 66 enthält einer Liste seiner Aufgaben; weitere Aufgaben werden im AIA bestimmt). Dazu unterstützt es das AI Office u.a. bei der Erstellung von Praxisleitfäden. Bei der Anwendung der DSGVO vertreten der EDPB und die Kommission oft gegensätzliche Positionen; ob dies auch beim AIA so sein wird, bleibt abzuwarten.

54 Welche weiteren EU-Stellen sieht der AIA vor?

Ein **Beratungsforum** unterstützt das EAIB und die Kommission mit technischem Fachwissen. Das Beratungsforum setzt sich aus Vertretern der Industrie, Start-Ups, KMU, der Zivilgesellschaft und der Wissenschaft sowie Europäischer Einrichtungen (z.B. Europäisches Komitee für Normung CEN oder der ENISA) zusammen (Art. 67).

Daneben soll die Kommission ein wissenschaftliches Gremium unabhängiger Sachverständiger (**wissenschaftliches Gremium**; "Scientific panel of independent experts") bilden. Es setzt sich aus unabhängigen Sachverständigen zusammen und soll das AI Office bei seinen Marktaufsichtstätigkeiten mit wissenschaftlichem und technischem Fachwissen unterstützen (Art. 68).

55 Welche Rolle haben die nationalen Marktüberwachungsbehörden?

Die **Marktüberwachungsbehörden** (Art. 3 Nr. 26 und 48) sind für die Marktüberwachung bei HRAIS und GPAIM zuständig (Art. 74 ff.). Jeder Mitgliedstaat muss mindestens eine solche Behörde bestellen (Art. 70 Abs. 1). Bei regulierten Produkten (Art. 6 Abs. 1) sind i.d.R. die dort zuständigen Behörden auch die Marktüberwachungsbehörde nach dem AIA (Art. 74 Abs. 3), im Finanzbereich die Finanzmarktaufsicht (Art. 74 Abs. 6), und für öffentliche Stellen der EU ist es der EDPS (Art. 74 Abs. 9). Ein Sonderfall sind AIS, die auf einem selbstentwickelten GPAIM beruhen (z.B. ChatGPT); hier liegt die Marktüberwachung beim AI Office (→ 52).

Die **Befugnisse und Aufgaben** richten sich dabei insbesondere nach der Marktüberwachungsverordnung (Art. 3 Nr. 26; Art. 14 ff. dieser Verordnung; https://dtn.re/QCMYaE) und den Anforderungen von Art. 70 Abs. 1. Beispielsweise können sie

- im Fall eines **schwerwiegenden Vorfalls** bei einem HRAIS anordnen, dass es zurückgerufen oder vom Markt genommen oder dass seine Bereitstellung auf dem Markt unterbleibt (Art. 19 der Marktüberwachungsverordnung);
- für ihre Tätigkeit jederzeit Informationen der Anbieter anfordern (Art. 74 Abs. 12 und 13 und 75 Abs. 3), die **Kooperationspflichten** haben, und sie können
- u.U. können sie einen **Test von HRAIS** anordnen (Art. 77 Abs. 3).
- Sollte ein AIS ein **besonderes Risiko** für die Gesundheit und Sicherheit von Personen, die Gesundheit und Sicherheit am Arbeitsplatz, den Verbraucherschutz, die Umwelt, die öffentliche Sicherheit und andere öffentliche Interessen darstellen (Art. 79 Abs. 1 i.V.m. Art. 3 Nr. 19 der Marktüberwachungsverordnung), kann die zuständige Marktüberwachungsbehörde die Konformität des betreffenden AIS prüfen und ggf. Korrekturmassnahmen anordnen und einen Rückruf anordnen (Art. 79 Abs. 5).
- Bei AIS, die der Anbieter als **nicht hochriskant** eingestuft hat, kann die Marktüberwachungsbehörde – wenn sie anderer Ansicht ist – anordnen, dass die Konformität hergestellt wird (Art. 80 Abs. 1 und 2). Sie kann auch bei zwar konformen, aber dennoch **besonders riskanten HRAIS** Abhilfemassnahmen anordnen (Art. 82 Abs. 1). Massnahmen kann sie ferner bei formalen Fehlern ergreifen, bspw. wenn ein CE-Kennzeichen fehlt (Art. 83).

Die Marktüberwachungsbehörden haben nach dem AIA ferner insbesondere die folgenden Aufgaben:

- **Information der Kommission** über bestimmte Rechtsvorschriften im Zusammenhang mit der biometrischen Echtzeit-Fernidentifizierung in zu Strafverfolgungszwecken (Art. 5 Abs. 4 und 6);
- **Entgegennahme von Informationen** und Meldungen, insbesondere der folgenden:
  - der Anbieter und Betreiber von HRAIS über besondere Risiken (Art. 79 Abs. 2 und Art. 26 Abs. 5); der Betreiber eines HRAIS über schwerwiegende Vorfälle (Art. 26 Abs. 5);
  - Kopie des Bestellungsauftrags und seiner Beendigung von Vertretern aussereuropäischer HRAIS-Anbieter (Art. 22 Abs. 3 und 4);
  - Information über nicht-konforme HRAIS von Einführern (Art. 23 Abs. 2);
  - Grundrechte-Folgenabschätzungen (FRIA) der öffentlichen Stellen (Art. 27 Abs. 3);
  - Informationen über Tests von HRAIS unter Realbedingungen (Art. 60);
  - Meldungen über schwerwiegende Vorfälle bei HRAIS (Art. 73 Abs. 1);
- **Information weiterer Stellen**:

- nationaler Behörden und öffentliche Stellen nach Art. 77, wenn ihr ein schwerwiegender Vorfall bei einem HRAIS gemeldet wurde (Art. 73 Abs. 7);
- der Kommission bei Massnahmen im Fall schwerwiegender Vorfälle (Art. 19 Abs. 1 der Marktüberwachungsverordnung);
- jährliche Berichterstattung z.Hd. der Kommission (Art. 74 Abs. 2);
- **ausnahmsweise Zulassung** eines HRAIS nach Art. 46;
- Genehmigung und Überprüfung der **Tests von HRAIS unter Realbedingungen** (Art. 60 Abs. 4 lit. b, Art. 76); ggf. auch Einschreiten, wenn es zu einem schwerwiegenden Vorfall kommt oder ein Test nicht den anwendbaren Bedingungen entsprechend verläuft (Art. 76 Abs. 3 und 5);
- Entgegennahme von **Beschwerden** natürlicher oder juristischer Personen (Art. 85).

56 Welche Rolle haben die Konformitätsbewertungsstellen?

Konformitätsbewertungsstellen führen die **Konformitätsbewertung** durch (Art. 3 Nr. 21). Sie werden von den notifizierenden Behörden (Art. 28 Abs. 1, 29 Abs. 1 und 30 Abs. 1 → 57) eingesetzt und müssen die Anforderungen nach Art. 30 erfüllen. Insbesondere müssen sie unabhängig sein. Auch Konformitätsbewertungsstellen in Drittstaaten können nach dem AIA tätig werden, sofern mit der EU ein entsprechendes Abkommen besteht (Art. 39). Eine Konformitätsstelle heisst "**notifizierte Stelle**", wenn sie nach den einschlägigen Bestimmungen notifiziert wurde (Art. 3 Nr. 22). Zum Konformitätsbewertungsverfahren → 15.

57 Welche Rolle haben die notifizierenden Behörden?

Jeder Mitgliedstaat muss eine notifizierende Behörde bestellen (Art. 28 Abs. 1 und 70 Abs. 1). Sie ist dafür zuständig, die Verfahren für die Bewertung, Benennung und Notifizierung von **Konformitätsbewertungsstellen einzurichten** und durchzuführen und diese zu überwachen (Art. 3 Nr. 19). "Notifizierenden Behörden" heissen sie, weil sie die Kommission (→ 51) und die übrigen Mitgliedstaaten über ein von der Kommission verwalteten Notifizierungsinstrument über jede Konformitätsbewertungsstelle informieren müssen; erst dadurch werden die Konformitätsbewertungsstellen zu notifizierten Stellen und können ihre Arbeit aufnehmen (→ 56).

Ergänzungsfragen

58 Welche Rolle spielt der Datenschutz im AIA?

Der Datenschutz hat eine erhebliche Bedeutung bei AIS, insbesondere im Zusammenhang mit dem Training von GPAIM. Der AIA verweist deshalb des Öfteren auf die DSGVO, insbesondere für dort legaldefinierte Begriffe (Art. 3 Nr. 37, 50, 51 und 52) oder deklaratorisch auf Vorgaben der DSGVO (bspw. in Art. 26 Abs. 9 für die Verwendung der Betriebsanleitung bei einer Datenschutz-Folgenabschätzung oder in Art. 50 Abs. 3 für die Information Betroffener), stellt aber klar, dass die DSGVO bei der Bearbeitung von Personendaten uneingeschränkt gilt (Art. 2 Abs. 7, 10 Abs. 5).

Art. 10 Abs. 5 enthält die einzige besondere **Rechtsgrundlage** im AIA. Zwischen der Datensparsamkeit und der Relevanz der Trainingsdaten besteht ein Zielkonflikt. Der AIA löst ihn so, dass ausnahmsweise sogar besonders schützenswerte Personendaten verarbeitet werden dürfen, wenn dies beim Training eines HRAIS unbedingt notwendig ist, um Bias zu erkennen und zu reduzieren (genauer: das Verbot von Art. 9 Abs. 1 DSGVO ist insoweit aufgeboben; eine Rechtsgrundlage nach Art. 6 bleibt erforderlich; EuGH, Rs. C-667/21, https://dtn.re/ATzHFf). Einzuhalten sind aber die besonderen Bedingungen nach Art. 10 Abs. 5 lit. a-f.

Wichtiger als diese Frage ist die Diskussion um den Anwendungsbereich des Datenschutzrechts auf ein Training von LLM und breiter, wann **im Rahmen eines LLM Personendaten** verwendet werden, welche Partei dabei welche Rolle einnimmt und wie Betroffenenrechte sichergestellt werden können. Dazu findet derzeit eine Diskussion statt. Zu verweisen ist insbesondere auf folgende Dokumente und Stellungnahmen (chronologisch absteigend):

- **David Rosenthal**, Blogbeitrag, 17. Juli 2024 (https://dtn.re/0SBhz2)
- **David Vasella,** datenrecht.ch, 16. Juli 2024 (https://dtn.re/BuTaCE)
- **Hamburg**: Diskussionspapier “Large Language Models und personenbezogene Daten” des HmbBfDI, 15. Juli 2024 (https://dtn.re/BuTaCE)

Die namentlich vom HmbBfDI vertretene Haltung, ein LLM könne keine Personendaten enthalten, weil es Inputdaten nicht kopiert, sondern Beziehungen zwischen Tokens als Vektoren bzw. Tensoren mathematisch darstellt, greift dabei zu kurz, weil der Aggregatszustand eines Personendatums nicht relevant ist: Werden personenbezogene Informationen nicht als solche, sondern in Form mathematischer Beziehungen in grundsätzlich wiedergabefähiger Form gespeichert, handelt es sich dabei um eine Bearbeitung von Personendaten (vgl. datenrecht.ch, https://dtn.re/BuTaCE). Die Frage, wie bspw. Betroffenenrechte bei LLMs umgesetzt werden können, ist daher nicht gegenstandslos.

Datenschutzbehörden haben sich auch ausserhalb der Frage des Personenbezugs von Embeddings (→ 12) zum **Verhältnis zwischen Datenschutz und künstlicher Intelligenz** geäussert, bspw.:

- **EDSA**, Statement 3/2024 on data protection authorities’ role in the Artificial Intelligence Act framework, 16. Juli 2024 (https://dtn.re/vGUUWh)
- **DSK**, Orientierungshilfe Künstliche Intelligenz und Datenschutz, 6. Mai 2024 (https://dtn.re/S63kDn)
- **BayLDA**, im 29. Tätigkeitsbericht 2019 (https://dtn.re/rg7FEr)
- **ICO**, verschiedene Informationen zu AI-Themen (https://dtn.re/g91v0E)
- **Österreich**: FAQ zum Thema KI und Datenschutz der österreichischen DSB, 2. Juli 2024 (https://dtn.re/Sz4sDS)

- **Frankreich**: CNIL, Self-assessment guide for artificial intelligence (AI) systems (https://dtn.re/44hM5n)
- **Italien**: Garante, Hinweise zum Schutz von Personendaten vor Scraping, 20. März 2024 (https://dtn.re/TuzT85)
- **Schweiz**: Siehe → 63

Mehrere europäische Datenschutzaufsichtsbehörden ("SAs") hatten ferner **Untersuchungen gegen OpenAI** im Zusammenhang mit ChatGPT eingeleitet. Der EDSA hat im April 2023 eine entsprechende Taskforce eingerichtet, deren Arbeit noch läuft; am 23. Mai 2024 wurde ein knapper Zwischenbericht veröffentlicht (https://dtn.re/HyvPHo).

59 Wie geht der AI Act mit Urheberrechten um?

Im Bereich des Urheberrechts erkennt der AIA das Problem des Trainings mit geschützten Werken. Er äussert sich zu dieser Problematik nicht inhaltlich, verlangt aber von den Anbietern von **GPAIM** u.a., über eine Strategie zur Einhaltung des EU-Urheberrechts zu verfügen und eine Zusammenfassung der Trainingsdaten zu veröffentlichen (→ 39).

Ansonsten bleibt die Zuweisung von Ausschliesslichkeitsrechten und die Bestimmung ihres Umfangs und entsprechenden Schranken aber den einschlägigen Bestimmungen überlassen. In diesem Zusammenhang wird vor allem diskutiert, unter welchen Voraussetzungen die Verwendung urheberrechtlich geschützter Werke für das **Training eines LLM** rechtsverletzend ist – eine verständliche Diskussion, nachdem LLMs insbesondere auch mit den Kreativen konkurrieren, mit deren Werken sie trainiert wurden.

Dabei gilt der **Territorialitätsgrundsatz**: Ob eine Handlung urheberrechtsverletzend ist, bestimmt das Recht des Staates, in dem sie zu verorten ist (für das schweizerische Kollisionsrecht: Art. 110 IRPG). In der EU sind das die Urheberrechte der einzelnen Mitgliedstaaten. Siehe aber → 40 zur Frage, ob die Anforderung an GPAIM-Anbieter ausserhalb der EU die Vorgaben des AIA zur Urheberrechtsstrategie einzuhalten haben.

In der Schweiz finden sich die einschlägigen Regelungen im URG. Offen ist insbesondere die Reichweite der Schrankenbestimmungen, wobei zu unterscheiden ist zwischen einer Beschaffung urheberrechtlich geschützten Materials und seiner Verwendung für ein Training:

Die Beschaffung und eine damit in aller Regel einhergehende **Vervielfältigung** von Material ist im Gegensatz zum Werkgenuss wie bspw. dem Durchsuchen von Text oder dem Labeling für ein Supervised Learning (→ 10) urheberrechtlich relevant (solange man den Vervielfältigungsbegriff nicht auf Handlungen beschränken will, die eine Wahrnehmbarmachung des Werks bezwecken). Fehlt eine **Lizenz** – die ausdrücklich oder stillschweigend eingeräumt werden kann –, stellt sich daher die Frage, ob die Schranke des Eigengebrauchs nach Art. 19 Abs. 1 URG greift. Hier besteht derzeit **Rechtsunsicherheit**:

- Es wird u.a. diskutiert, ob ein Training unter die Vervielfältigung und Bereitstellung für die "**interne Information oder Dokumentation**" fallen kann (Art. 19 Abs. 1 lit. c URG). Da eine solche Vervielfältigung im Wesentlichen nur für nicht-kommerzielle Zwecke freigestellt ist und das Training eines LLMs daher in der Regel nicht erfassen dürfte und da die Vervielfältigung im Handel erhältlicher Werkexemplare nicht abdeckt ist (Art. 19 Abs. 3 lit. a URG), wird diese Schranke häufig nicht greifen.
- Ebenfalls diskutiert wird sog. "**Text-and-Data-Mining**" (TDM), das eine Vervielfältigung zu wissenschaftlichen Zwecke freistellt, wenn sie technisch bedingt ist, bspw. durch die semantische Analyse des Ausgangsmaterials (Art. 24d URG). Der Wissenschaftsbegriff ist zwar breit, doch verlangt auch die angewandte Forschung privater Unternehmen einen ernsthaften Erkenntniszweck. Ob der Umstand, dass ein trainiertes LLM für unterschiedliche Zwecke verwendbar ist, genügt, dem Training den erforderlichen Erkenntniszweck zuzusprechen, ist ungewiss; jedenfalls reicht es nicht,

dass ein trainiertes LLM ggf. zu Forschungszwecken verwendet werden kann, der Forschungszweck müsste das Training erfassen.

Ausserdem muss die Beschaffung der verwendeten Werke rechtmässig sein (Art. 24d URG), was u.a. bei öffentlich verfügbaren Werken nicht pauschal bejaht (oder verneint werden kann).

– Eine nur **flüchtige Vervielfältigung** wäre zwar freigestellt, solange sie nur flüchtig oder begleitend ist, einen integralen und wesentlichen Teil eines technischen Verfahrens darstellt, ausschliesslich der Übertragung in einem Netz zwischen Dritten durch einen Vermittler oder einer rechtmässigen Nutzung dient und keine eigenständige wirtschaftliche Bedeutung hat (Art. 24a URG). Diese Voraussetzungen – sie gelten kumulativ – dürften auf die Zusammenstellung eines Trainings-, Tests- und/oder Validierungsdatensatzes kaum zutreffen (und wohl auch kaum auf den Trainingsvorgang selbst, der eine erhebliche wirtschaftliche Bedeutung hat). Auch Art. 24a URG ist daher kaum eine Grundlage für den gesamten Trainingsvorgang mit urheberrechtlich geschütztem Material.

Der Output seinerseits ist urheberrechtlich kaum geschützt, weil eine geistige, d.h. **menschliche Schöpfung** fehlt (Art. 2 Abs. 1 URG); dies jedenfalls, soweit der Output nicht nachweislich von einer natürlichen Person vorgegeben wurde. Eine AI kann aus dem gleichen Grund kein Erfinder im patentrechtlichen Sinne sein – auch hier setzt der Schutz voraus, dass der Erfinder ein Mensch ist.

60 Was gilt beim Einsatz von AI am Arbeitsplatz?

Der AIA enthält einige wenige Bestimmungen eigens im Zusammenhang mit dem Einsatz von AIS im Arbeitskontext:

– Der Einsatz eines AIS ist nach Art. 5 in wenigen Fällen **verboten** (→ 27). Das kann im Arbeitsbereich der Fall sein, bspw. bei Emotionserkennung am Arbeitsplatz, wenn die Schutzbedürftigkeit von Mitarbeitenden ausgenutzt werden soll oder wenn ein Social Scoring erfolgen würde;

– Der Begriff des HRAIS umfasst **arbeitsplatzbezogene Use Cases** (→ 28), bspw. wenn AIS zur Steuerung des Zugangs zur beruflichen Aus- und Weiterbildung verwendet wird, bei der Einstellung oder Auswahl von Stellenbewerbungen zum Einsatz kommt oder bei Entscheidungen über Arbeitsbedingungen, Beförderungen oder Kündigungen verwendet wird.

– Vor der Inbetriebnahme oder Verwendung eines HRAIS am Arbeitsplatz muss der Betreiber die **Mitarbeitervertreter** und die betroffenen Arbeitnehmenden informieren, dass sie "der Verwendung des Hochrisiko-KI-Systems unterliegen" werden (Art. 26 Abs. 7 AIA).

– Zu informieren ist auch dann, wenn ein HRAIS verwendet wird – auch, aber nicht nur im Arbeitskontext –, um **Entscheidungen** zu treffen oder dabei zu unterstützen (Art. 26 Abs. 11 AIA).

Ansonsten bleibt der Schutz der Arbeitnehmenden und Bewerbenden aber den sonstigen Vorschriften des **anwendbaren Rechts** überlassen, insbesondere des Datenschutzrechts und des öffentlichen Arbeitsrechts, das Mitwirkungsrechte vorsehen kann.

In der EU sind allerdings Gesetzgebungsprojekte im Gang, die den Schutz der Arbeitnehmenden verbessern sollen. Die im Entwurf vorliegende Plattformrichtlinie der EU (https://dtn.re/ttWVev) enthält im Kapitel III besondere Regeln für das "algorithmische Management" im Bereich der Plattformarbeit. Die Plattformrichtlinie muss von den Mitgliedstaaten umgesetzt werden, wenn sie verabschiedet wird – sie ist eine Richtlinie und anders als die DSGVO oder der AIA keine direkt anwendbare Verordnung. Das Parlament hat die Plattformrichtlinie am 24. April 2024 verabschiedet (https://dtn.re/G3ytlM), die Zustimmung des Rats steht noch aus.

61 Welche internationalen Standards betreffen AI?

Mehrere Standards und Normungsinitiativen befassen sich mit AI. Die International Organization for Standardization (**ISO**) und die International Electrotechnical Commission (**IEC**) haben gemeinsam Standards entwickelt:

- ISO/IEC 42001:2023 (https://dtn.re/L8KOIs): Anforderungen an AI-Managementsysteme
- ISO/IEC TR 24028:2020 (https://dtn.re/YYy0Ha): Vertrauenswürdigkeit von AI-Systemen, Kriterien für Transparenz, Kontrolle und Erklärbarkeit
- ISO/IEC 5259-1: Basis der ISO 5259-Reihe betr. Datenqualität für Analysen und ML (https://dtn.re/Tggl5G)
- ISO/IEC TR 5469:2024: Einsatz von AI in sicherheitsrelevanten Funktionen (https://dtn.re/vbc8IL)

In Europa beteiligen sich das CEN (Europäisches Komitee für Normung) und das CENELEC (Europäisches Komitee für elektrotechnische Normung) an der Entwicklung von AI-Standards über das gemeinsame Komitee **CEN-CENELEC JTC 21 "Artificial Intelligence"**. Es hat mehrere Standards veröffentlicht, und weitere befinden sich in Ausarbeitung (https://dtn.re/Gx0XMT). Veröffentlicht wurden bspw.:

- CEN/CLC ISO/IEC/TR 24027:2023: Bias ()
- CEN/CLC ISO/IEC/TR 24029-1:2023: Beurteilung der Robustheit neuronaler Netzwerke

Das amerikanische National Institute of Standards and Technology (**NIST**) hat sodann ein AI-Risikomanagement-Framework entwickelt, das AI RMF 1.0, veröffentlicht im Januar 2023, das anschliessend durch "Profiles" ergänzt wurde, Implementierungen für bestimmte Umstände, Anwendungen oder Technologien. Ein Beispiel ist NIST AI 600-1 "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile" (https://dtn.re/z3H7BJ).

62 Was ist die AI-Konvention des Europarats?

Der Europarat (nicht der Rat der Europäischen Union) hat am 17. Mai 2024 die AI-Konvention des Europarats (Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, AI Convention) verabschiedet. Der Text der AI Convention ist zusammen mit dem Explanatory Report auf datenrecht.ch verfügbar (https://dtn.re/8zndsz, auf Englisch).

Die Konvention ist eine von den ratifizierenden Staaten – zu der die Schweiz sicher gehören wird – umzusetzende Rahmenvereinbarung, die Standards in Bezug auf Menschenrechte, Demokratie und Rechtsstaatlichkeit beim Einsatz von AI-Systemen sicherstellen soll.

Mitglieder und Nichtmitglieder des Europarates werden nun aufgefordert, das Rahmenübereinkommen zu unterzeichnen und zu ratifizieren. Sollte die Schweiz die Konvention ratifizieren, muss sie sie ins schweizerische Recht überführen (→ 63).

Die Vorgaben der AI Convention sind sehr vage. Dazu kommt, dass sie die Mitgliedstaaten nur bei der Legiferierung im öffentlichen Sektor bindet. Im Privatbereich sollen die Mitgliedstaaten lediglich in einer Weise, die "mit Ziel und Zweck" der AI Convention "vereinbar ist" (Art. 3 Abs. 1 der AI Convention).

63 Wie reguliert die Schweiz den Einsatz künstlicher Intelligenz?

In der Schweiz existiert bisher keine übergreifende Regulierung des Einsatzes künstlicher Intelligenz. Der Bundesrat hat das UVEK Ende 2023 beauftragt, im Rahmen der Interdepartementalen Koordinationsgruppe EU-Digitalpolitik **bis Ende 2024 mögliche Ansätze** für eine Regulierung auszuloten (siehe die Medienmitteilung, https://dtn.re/uV1Eau). Das UVEK bzw. in dessen Auftrag das BAKOM soll dabei vom geltenden Recht ausgehen und Regulierungsansätze finden, die sowohl mit dem AIA als auch der AI Convention (→ 62) kompatibel sind.

Bis Ende 2024 sollen die Auslegeordnung des BAKOM einschliesslich der dafür angefertigten Grundlagenstudien bspw. zu Regelungslücken des geltenden Rechts sowie der Richtungsentscheid des Bundesrats vorliegen.

Welche Ansätze das UVEK vorschlägt und welche sich letztlich durchsetzen, ist derzeit allerdings offen. Eine **Vollübernahme des AIA** dürfte politisch wenig Chancen haben, solange die EU dies nicht als Bedingung für die Teilnahme am Binnenmarkt stellt, und die AI Convention ist so vage, dass ihr Inhalt eine Regulierung kaum vorzeichnet, besonders nicht im privaten Bereich (→ 62). Die Wirtschaft (aber auch die Academia) pocht auf schlanke Regelungen, während zivilgesellschaftliche Organisationen strengere Bestimmungen insbesondere zum Schutz vor Diskriminierung fordern (z.B. Algorithm Watch). Am naheliegendsten erscheint derzeit ein Mantelerlass, der die einschlägigen Rechtsgrundlagen punktuell anpasst.

Zudem sind diverse politische Vorstösse hängig, bspw. die folgenden (auf Bundesebene):

- 24.3796, Motion **Glättli**, 14. Juni 2024, Transparente risikobasierte Folgeabschätzungen bei Einsatz von KI und Algorithmen durch den Bund (https://dtn.re/vWwoDP)
- 24.3795, Motion **Glättli**, 14. Juni 2024, Schutz vor Diskriminierung beim Einsatz von KI und Algorithmen (https://dtn.re/B46Qtc)
- 24.3611, Interpellation **Cottier**, 13. Juni 2024, Künstliche Intelligenz. Koordination in der Verwaltung und Absichten bezüglich des neuen Rahmenübereinkommens des Europarats (https://dtn.re/hdDPxQ)
- 24.3616, Interpellation **Gössi**, 13. Juni 2024, Medien und künstliche Intelligenz (https://dtn.re/JaEh4n)
- 24.3415, Interpellation **Tschopp**, 17. April 2024, Plattformen und KI: Rechte der Nutzerinnen und Nutzer (https://dtn.re/HBZFOE)
- 24.3363, Motion **Chappuis**, 15. März 2024, Für eine souveräne digitale Infrastruktur in der Schweiz im Zeitalter der künstlichen Intelligenz (https://dtn.re/s4SsC9)
- 24.3346, Interpellation **Docourt**, 15. März 2024, EU-Richtlinie über Plattformarbeit. Will sich die Schweiz daran orientieren? (https://dtn.re/UNvBOq)
- 24.3235, Interpellation **Marti**, 14. März 2024, Künstliche Intelligenz und die Auswirkungen auf das Urheberrecht (https://dtn.re/jpX0Cg)
- 24.3209, Motion **Juillard**, 14. März 2024, Für eine souveräne digitale Infrastruktur in der Schweiz im Zeitalter der künstlichen Intelligenz (KI) (https://dtn.re/NsqdKN)
- 23.4517, Interpellation **Gugger**, 22. Dezember 2023, Künstliche Intelligenz und Mitwirkung. Gibt es Lücken im Gesetz? (https://dtn.re/hl1Q54)
- 23.4492, Motion **Gysi**, 22. Dezember 2023, Künstliche Intelligenz am Arbeitsplatz. Mitwirkungsrechte der Arbeitnehmenden stärken (https://dtn.re/PH8ab1)
- 23.4051, Interpellation **Schlatter**, 29. September 2023, Künstliche Intelligenz und Robotik. Ethik gehört in die Ausbildung! (https://dtn.re/PMNgtC)
- 23.393, Interpellation **Cottier**, 16. Juni 2023, Künstliche Intelligenz. Welche Rahmenbedingungen müssen geschaffen werden, um das Beste daraus zu machen und Fehlentwicklungen zu vermeiden? (https://dtn.re/FXxB9v)

- 23.3812, Interpellation **Widmer**, 15. Juni 2023, Künstliche Intelligenz. Gefahren und Potenziale für die Demokratie (https://dtn.re/ZkaTUc)
- 23.4133, Interpellation **Marti**, 28. September 2023, Algorithmische Diskriminierung. Ist der gesetzliche Diskriminierungsschutz ausreichend? (https://dtn.re/xr97Zq)
- 23.3849, Motion **Bendahan**, 15. Juni 2023, Ein Kompetenzzentrum oder Kompetenznetzwerk für künstliche Intelligenz in der Schweiz schaffen (https://dtn.re/sqLWYa)
- 23.3654, Interpellation **Riniker**, 13. Juni 2023, Rolle der Schweiz in der internationalen Zusammenarbeit auf dem Gebiet der künstlichen Intelligenz (https://dtn.re/sUoUb3)
- 23.3806, Motion **Marti**, 15. Juni 2023, Deklarationspflicht bei Anwendungen der künstlichen Intelligenz und automatisierten Entscheidungssystemen (https://dtn.re/D3FmNo)
- 23.3563, Motion **Mahaim**, 4. Mai 2023, Deepfakes regulieren (https://dtn.re/kwNWvh)
- 23.3516, Interpellation **Feller**, 2. Mai 2023, Grundsätzliches oder vorläufiges Verbot von bestimmten Plattformen der künstlichen Intelligenz (https://dtn.re/Ig8JPJ)
- 23.3201, Postulat **Dobler**, 16. März 2023, Rechtslage der künstlichen Intelligenz. Unsicherheiten klären, Innovation fördern! (https://dtn.re/e7sGIM)
- 23.3147, Interpellation **Bendahan**, 14. März 2023, Regulierung der künstlichen Intelligenz in der Schweiz (https://dtn.re/xMVLIE)
- 21.4406, Postulat **Marti**, 9. Dezember 2021, Bericht zur Regulierung von automatisierten Entscheidungssystemen (https://dtn.re/PQbXqs)
- 21.3206, Interpellation **Pointet**, 17. März 2021, Welche Prozesse des Staates stützen sich auf künstliche Intelligenz? (https://dtn.re/WUw9Hr)
- 21.3012, Postulat **Sicherheitspolitische Kommission**, 15. Januar 2021, Klare Regeln für autonome Waffen und künstliche Intelligenz (https://dtn.re/duRhvk)
- 19.3919, Interpellation **Riklin**, 21. Juni 2019, Künstliche Intelligenz und digitale Transformation. Wir brauchen eine holistische Strategie (https://dtn.re/5x93tL)

Selbstverständlich gelten die sonst anwendbaren Bestimmungen auch beim Einsatz von KI. Das betrifft bspw.

- das **Datenschutzrecht** (wenn beim Training oder beim Einsatz Personendaten bearbeitet werden),
- das **Geheimnisschutzrecht** (wenn geheime Informationen für ein Training oder als Input verwendet werden),
- das **Arbeitsvertragsrecht** (wenn Personendaten von Bewerbenden und Mitarbeitenden bearbeitet werden und wenn eine KI die Fürsorgepflicht des Arbeitgebers tangiert),
- das **öffentliche Arbeitsrecht** (z.B. wenn Mitwirkungspflichten greifen oder eine Verhaltensüberwachung zur Diskussion steht),
- das **Persönlichkeitsrecht** (z.B. wenn Gespräche oder Teamscalls aufgezeichnet werden),
- das **Lauterkeitsrecht** (wenn AI-generierte Inhalte irreführend sein können),
- das **Urheberrecht** (z.B. wenn eine AI mit Werken trainiert oder Werke als Input verwendet werden, und wenn der Schutz von Output zur Diskussion steht),
- das **Strafrecht** (bei Aufnahmen nicht-öffentlicher Gespräche oder generell beim Einsatz von AI bei strafbarem Verhalten),
- weitere Rechtsgebiete.

Auch sektorielle Regelungen können betroffen sein. Wenige Aufsichtsbehörden haben bereits Erwartungen formuliert, so insbesondere die **FINMA** (https://dtn.re/HiLkOo). Andere Behörden warten offenbar noch zu (bspw. das Bundesamt für Gesundheit). In der Bundesverwaltung selbst gelten seit 2020 die “Leitlinien ‘Künstliche Intelligenz’ für den Bund” (https://dtn.re/VdyAxo). Auf kantonaler Ebene bestehen ebenfalls Vorgaben oder Empfehlungen, bspw. zu “Rechtliche Best Practices” beim Einsatz von AI in der Bildung im Kanton Zürich (https://dtn.re/bOT1Ez).

Private Akteure haben sich in der Zwischenzeit ebenfalls Regeln gegeben. Das betrifft vor allem besonders exponierte Akteure wie

- Medien (die Publizistischen Leitlinien der SRG (https://dtn.re/AgauQE) sind mit einer besonderen Regelung des Einsatzes von AI ein Beispiel, ein weiteres ist der Leitfaden Künstliche Intelligenz des Presserats, (https://dtn.re/f1UTYZ),
- politische Parteien (bspw. mit dem KI-Kodex der Grünen, der GLP, der SP, der Mitte und der EVP, https://dtn.re/0b4WvK, oder der Selbstverpflichtung der FDP, https://dtn.re/1te4U8) oder
- die Forschung und Ausbildung (bspw. mit den Empfehlungen zum Umgang mit generativer Künstlicher Intelligenz an der UZH, https://dtn.re/aBstLV).

Auch zahlreiche private Unternehmen haben teils öffentliche, teils nicht öffentliche Richtlinien, Kodizes und Anweisungen erlassen oder sind dabei, es zu tun.

Anhang: In Art. 3 AIA definierte Begriffe

	Englisch	Deutsch
1	AI system	KI-System
2	Risk	Risiko
3	Provider	Anbieter
4	Deployer	Betreiber
5	Authorised representative	Bevollmächtigter
6	Importer	Einführer
7	Distributor	Händler
8	Operator	Akteur
9	Placing on the market	Inverkehrbringen
10	Making available on the market	Bereitstellung auf dem Markt
11	Putting into service	Inbetriebnahme
12	Intended purpose	Zweckbestimmung
13	Reasonably foreseeable misuse	Vernünftigerweise vorhersehbare Fehlanwendung
14	Safety component	Sicherheitsbauteil
15	Instructions for use	Betriebsanleitungen
16	Recall of an AI system	Rückruf eines KI-Systems
17	Withdrawal of an AI system	Rücknahme eines KI-Systems
18	Performance of an AI system	Leistung eines KI-Systems
19	Notifying authority	Notifizierende Behörde
20	Conformity assessment	Konformitätsbewertung
21	Conformity assessment body	Konformitätsbewertungsstelle
22	Notified body	Notifizierte Stelle

23	Substantial modification	Wesentliche Veränderung
24	CE marking	CE-Kennzeichnung
25	Post-market monitoring system	System zur Beobachtung nach dem Inverkehrbringen
26	Market surveillance authority	Marktüberwachungsbehörde
27	Harmonised standard	Harmonisierte Norm
28	Common specification	Gemeinsame Spezifikation
29	Training data	Trainingsdaten
30	Validation data	Validierungsdaten
31	Validation data set	Validierungsdatensatz
32	Testing data	Testdaten
33	Input data	Eingabedaten
34	Biometric data	Biometrische Daten
35	Biometric identification	Biometrische Identifizierung
36	Biometric verification	Biometrische Verifizierung
37	Special categories of personal data	Besondere Kategorien personenbezogener Daten
38	Sensitive operational data	Sensible operative Daten
39	Emotion recognition system	Emotionserkennungssystem
40	Biometric categorisation system	System zur biometrischen Kategorisierung
41	Remote biometric identification system	Biometrisches Fernidentifizierungssystem
42	Real-time remote biometric identification system	Biometrisches Echtzeit-Fernidentifizierungssystem
43	Post-remote biometric identification system	System zur nachträglichen biometrischen Fernidentifizierung
44	Publicly accessible space	Öffentlich zugänglicher Raum
45	Law enforcement authority	Strafverfolgungsbehörde

46	Law enforcement	Strafverfolgung
47	AI Office	Büro für Künstliche Intelligenz
48	National competent authority	Zuständige nationale Behörde
49	Serious incident	Schwerwiegender Vorfall
50	Personal data	Personenbezogene Daten
51	Non-personal data	Nicht personenbezogene Daten
52	Profiling	Profiling
53	Real-world testing plan	Plan für einen Test unter Realbedingungen
54	Sandbox plan	Plan für das Reallabor
55	AI regulatory sandbox	KI-Reallabor
56	AI literacy	KI-Kompetenz
57	Testing in real-world conditions	Test unter Realbedingungen
58	Subject	Testteilnehmer
59	Informed consent	Informierte Einwilligung
60	Deep fake	Deepfake
61	Widespread infringement	Weitverbreiteter Verstoss
62	Critical infrastructure	Kritische Infrastrukturen
63	General-purpose AI model	KI-Modell mit allgemeinem Verwendungszweck
64	High-impact capabilities	Fähigkeiten mit hoher Wirkkraft
65	Systemic risk	Systemisches Risiko
66	General-purpose AI system	KI-System mit allgemeinem Verwendungszweck
67	Floating-point operation	Gleitkommaoperation
68	Downstream provider	Nachgelagerter Anbieter