www.jusletter-it.eu

David Marti

Die KI-Konvention des Europarats: Ursprung, Inhalt, Ausblick

Der Beitrag erläutert die Entstehung, den Inhalt und die bevorstehenden Herausforderungen der Rahmenkonvention des Europarats zu KI und Menschenrechten, Demokratie und Rechtsstaatlichkeit, die im Mai 2024 angenommen wurde und das erste bindende internationale Regelwerk für KI-Systeme darstellt. Der Autor hebt die Notwendigkeit hervor, die Technologieunternehmen am Anfang der KI-Wertschöpfungskette in die Verantwortung zu nehmen und verweist auf ungelöste technische Herausforderungen, um die in der Konvention genannten Prinzipien in Bezug auf Tätigkeiten innerhalb des Lebenszyklus von KI-Systemen wirksam umzusetzen.

Beitragsart: Beiträge

Zitiervorschlag: David Marti, Die KI-Konvention des Europarats: Ursprung, Inhalt, Ausblick, in: Jusletter IT 4. Juli 2024

ISSN 1664-848X, jusletter-it.weblaw.ch, Weblaw AG, info@weblaw.ch, T +41 31 380 57 77

Inhaltsübersicht

1. Entstehungsgeschichte

[1] Die Rahmenkonvention des Europarats zu KI und Menschenrechten, Demokratie und Rechtsstaatlichkeit wurde am 17. Mai dieses Jahres vom Ministerrat des Europarats in Strassburg angenommen.[1] Der Prozess erstreckte sich über fünf Jahre: Als Reaktion auf die sich abzeichnenden gesellschaftlichen Herausforderungen von KI gründete der Ministerrat des Europarats am 11. September 2019 das Ad-hoc-Komitee für künstliche Intelligenz: CAHAI.[2]

[2] Die Hauptaufgabe des CAHAI lag darin, die Machbarkeit eines rechtlichen Rahmens für die Entwicklung, das Design und die Anwendung von KI zu prüfen, basierend auf den Standards des Europarats zu Menschenrechten, Demokratie und Rechtsstaatlichkeit. Dabei wurde ein Schwerpunkt gelegt auf den Einbezug verschiedener Stakeholder aus der Wissenschaft, der Zivilgesellschaft und der Wirtschaft.

[3] Mit dem Abschluss seines zweijährigen Mandats im Jahr 2021 empfahl das CAHAI die Erarbeitung einer KI-Konvention durch den Europarat. Dazu wurde das Committee on AI (CAI) einberufen, welches seine Arbeit unter Vorsitz des Schweizer Botschafters Thomas Schneider mit einer ersten Plenarsitzung im April 2022 aufnahm.[3]

[4] Das CAI inkludiert Delegationen der 46 Europarats-Mitgliedstaaten und nahm im Verlauf seiner Arbeiten die Beobachterstaaten Argentinien, Australien, Kanada, Costa Rica, den Vatikan, Israel, Japan, Mexiko, Peru, die USA und Uruguay auf. Zudem nahmen an den zehn Plenarsitzungen bis März 2024 teil: weitere Vertretungen des Europarats (darunter das Sekretariat der parlamentarischen Versammlung), Delegationen anderer internationaler Organisationen wie der

1 Council of Europe Treaty Series – No. [225], Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (https://rm.coe.int/1680afae3c, alle Websites zuletzt besucht am 17. Juni 2024).

2 Ad Hoc Committee on Artificial Intelligence – CAHAI (https://www.coe.int/en/web/artificial-intelligence/cahai-1#%7B%2266693401%22:[0]%7D).

3 Committee on Artificial Intelligence (CAI) (https://www.coe.int/en/web/artificial-intelligence/cai#%7B%22126720142%22:[1]%7D).

UNESCO, der OECD und der OSZE, Wirtschaftsvertreter:innen sowie Zivilgesellschaftsorganisationen (NGOs). Als Vertreter der NGOs hatte *Pour Demain* seit der sechsten Plenarsitzung offiziellen Beobachterstatus, nahm an den Plenarsitzungen teil und verfasste Textvorschläge für den Konventionsentwurf.

[5] Der Verhandlungsprozess glich während langer Strecken einem Balanceakt, da es der Anspruch der Konvention ist, auch für Staaten ausserhalb des Europarats kompatibel zu sein. Es war keine einfache Aufgabe, die teilweise recht unterschiedlichen Rechtstraditionen unter einen Hut zu bringen.

[6] Das CAI hatte als Ziel, einen möglichst inklusiven Verhandlungsprozess zu führen. Dies wurde aber mit der Einführung einer Drafting Group ab der dritten Plenarsitzung erschwert, zu der nur Länderdelegationen zugelassen waren. Das Sekretariat des CAI wies zwar darauf hin, dass Textvorschläge aller Delegationen nach wie vor gleichwertig behandelt würden, aber diese Entscheidung blieb bis zum Schluss der Verhandlungen ein Hauptkritikpunkt der NGOs.

[7] Der ursprüngliche Entwurf der Konvention erfuhr im Verlauf der zehn Plenarsitzungen diverse Änderungen, wobei Artikel 3 zum Geltungsbereich der Konvention der kontroverseste war. Die Kernfrage lautete: unter welchen Bedingungen liegen die Aktivitäten der Privatwirtschaft im Geltungsbereich der Konvention? Ein Kompromiss zu dieser Frage wurde erst spätabends am letzten Verhandlungstag erreicht und involvierte auch Formulierungen zum Schutz nationaler Sicherheitsinteressen sowie zu Forschungs- und Entwicklungsaktivitäten (ausser das infrage kommende KI-System hat das Potential, die Menschenrechte, die Demokratie und die Rechtsstaatlichkeit zu beeinträchtigen). Während den letzten zwei Plenarsitzungen wurde zudem der erläuternde Bericht verhandelt, in der Endversion 33 Seiten lang und für die Interpretation des Konventionstextes sehr relevant, da dieser auf zwölf Seiten sehr schlank gehalten ist.[4]

[8] Die finale Version der KI-Konvention wird am 5. September 2024 anlässlich der Konferenz der Justizminister der Europarats-Mitgliedstaaten in Vilnius zur Unterzeichnung aufgelegt. Danach finden bis Ende Jahr noch zwei Plenarsitzungen des CAI statt, um einen begleitenden Mechanismus zur Risiko- und Folgenabschätzung zu formulieren.

2. Inhalt der Konvention

[9] Die Rahmenkonvention über KI und Menschenrechte, Demokratie und Rechtsstaatlichkeit teilt sich über zwölf Seiten in eine Präambel und 36 Artikel auf. Die Artikel sind in acht Kapitel gegliedert:

- Kapitel I – Allgemeine Bestimmungen (Artikel 1–3)
- Kapitel II – Allgemeine Verpflichtungen (Artikel 4–5)
- Kapitel III – Grundsätze in Bezug auf Tätigkeiten innerhalb des Lebenszyklus von KI-Systemen (Artikel 6–13)
- Kapitel IV – Rechtsmittel (Artikel 14–15)

4 Council of Europe Treaty Series – No. [225], Explanatory Report to the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (https://rm.coe.int/1680afae67).

- Kapitel V – Bewertung und Minderung von Risiken und nachteiligen Auswirkungen (Artikel 16)
- Kapitel VI – Implementierung der Konvention (Artikel 17–22)
- Kapitel VII – Follow-up-Mechanismus und Zusammenarbeit (Artikel 23–26)
- Kapitel VIII – Schlussklauseln (Artikel 27–36)

2.1. Präambel

[10] Die Präambel betont das Ziel des Europarats, die Einheit seiner Mitglieder durch die Achtung der Menschenrechte, Demokratie und Rechtsstaatlichkeit zu fördern. Angesichts der rasanten wissenschaftlichen und technologischen Entwicklungen sowie der tiefgreifenden Veränderungen durch KI, die das menschliche Wohlergehen und die nachhaltige Entwicklung fördern können, wird betont, dass KI einzigartige Möglichkeiten bietet, Menschenrechte, Demokratie und Rechtsstaatlichkeit zu schützen und zu fördern.

[11] Es wird jedoch auch die Sorge geäussert, dass bestimmte KI-Aktivitäten die Menschenwürde und -rechte, Demokratie und Rechtsstaatlichkeit untergraben und Diskriminierung verstärken könnten. Insbesondere wird die Missbrauchsgefahr von KI-Systemen für repressive Zwecke, wie willkürliche Überwachung und Zensur, hervorgehoben.

[12] Die Präambel unterstreicht die Notwendigkeit eines globalen rechtlichen Rahmens, der gemeinsame Prinzipien und Regeln für KI-Aktivitäten festlegt und verantwortungsvolle Innovationen fördert. Zudem wird auf die Berücksichtigung weitergehender Risiken und Auswirkungen von KI, wie auf Gesundheit, Umwelt und sozioökonomische Aspekte, hingewiesen.

2.2. Kapitel I – Allgemeine Bestimmungen (Artikel 1–3)

[13] Mit Artikel 1 stellt die Konvention sicher, dass KI-Systeme mit Menschenrechten, Demokratie und Rechtsstaatlichkeit übereinstimmen. Vertragsparteien müssen geeignete Massnahmen zur Umsetzung ergreifen, abgestuft nach Umfang und Wahrscheinlichkeit negativer Auswirkungen. Die Vertragsparteien bekennen sich dazu, die Konvention weiterzuentwickeln durch internationale Zusammenarbeit.

[14] Artikel 2 definiert KI-Systeme und übernimmt fast wortgleich die aktualisierte Definition der OECD: Für die Zwecke dieser Konvention bedeutet «KI-System» ein maschinenbasiertes System, das für explizite oder implizite Ziele aus den ihm zugeführten Eingaben ableitet, wie es Ausgaben generieren kann, wie z.B. Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen, die physische oder virtuelle Umgebungen beeinflussen können. Verschiedene KI-Systeme variieren in ihren Autonomie- und Anpassungsstufen nach der Implementierung.

[15] Mit dem Geltungsbereich der Konvention wird in Artikel 3 das Kernstück des Textes dargelegt. Die Konvention gilt für Aktivitäten im gesamten Lebenszyklus von KI-Systemen, welche die Menschenrechte, Demokratie und Rechtsstaatlichkeit beeinträchtigen können, wie folgt:

a. Für Aktivitäten von öffentlichen Behörden oder privaten Akteuren im Auftrag der Behörden.
b. Für Aktivitäten von privaten Akteuren, soweit nicht unter Punkt a abgedeckt.

[16] Zu Punkt b: Jede Vertragspartei muss deklarieren, wie sie diese Verpflichtungen umsetzt, entweder durch Anwendung der Prinzipien der Konvention auf private Akteure oder durch andere Massnahmen. Änderungen an dieser Deklaration sind möglich. Internationale Verpflichtungen zum Schutz von Menschenrechten, Demokratie und Rechtsstaatlichkeit dürfen nicht eingeschränkt werden. Die Anwendung auf private Akteure wird somit jeder Vertragspartei selbst überlassen.

[17] Die Konvention gilt nicht für nationale Sicherheitsaktivitäten, sofern diese im Einklang mit internationalem Recht und demokratischen Prozessen stehen. Die Konvention gilt ebenso wenig für Forschung und Entwicklung von KI-Systemen, die noch nicht zur Nutzung bereitgestellt wurden, es sei denn, Testaktivitäten beeinträchtigen Menschenrechte, Demokratie und Rechtsstaatlichkeit. Angelegenheiten der nationalen Verteidigung fallen nicht unter die Konvention.

2.3. Kapitel II – Allgemeine Verpflichtungen (Artikel 4–5)

[18] Mit den allgemeinen Verpflichtungen wird festgehalten, dass die Menschenrechte zu schützen sind sowie die Integrität demokratischer Prozesse aufrechterhalten und der Respekt vor der Rechtsstaatlichkeit gewahrt wird – für Aktivitäten im gesamten Lebenszyklus von KI-Systemen. Spezifisch erwähnt werden unter anderem der Zugang zur Justiz, der faire Zugang zur öffentlichen Debatte sowie die Meinungsbildungsfreiheit.

2.4. Kapitel III – Prinzipien in Bezug auf Tätigkeiten innerhalb des Lebenszyklus von KI-Systemen (Artikel 6–13)

[19] Kapitel III beschäftigt sich mit allgemeinen Prinzipien beim Umgang mit KI-Systemen. Alle diese Artikel sind eher kurz gehalten. Entsprechend wichtig ist der erläuternde Bericht für die Interpretation dieser Artikel. Zuerst kommt ein Verweis darauf, dass diese Prinzipien entsprechend der nationalen Rechtssysteme von Vertragsparteien umgesetzt werden sollen. Der Respekt vor der menschlichen Würde und der individuellen Autonomie wird in Artikel 7 unterstrichen.

[20] Artikel 8 und 9 betreffen Transparenz-, Aufsichts- und Verantwortlichkeitsanforderungen: die Identifizierung von durch KI generierten Inhalten wird besonders erwähnt sowie Massnahmen zur Regelung von Haftungsfragen bei durch KI-Aktivitäten ausgelösten negativen Auswirkungen auf Menschenrechte, Demokratie und Rechtsstaatlichkeit.

[21] Artikel 10 und 11 behandeln Gleichheit und Nichtdiskriminierung sowie den Schutz personenbezogener Daten. Dabei wird Gendergleichheit herausgehoben sowie die Überwindung von Ungleichheiten, um gerechte und faire Ergebnisse zu erzielen. Beim Datenschutz wird die Gewährleistung wirksamer Garantien und Schutzmassnahmen eingefordert.

[22] Schliesslich behandeln Artikel 12 und 13 die Zuverlässigkeit von KI-Systemen sowie sichere Innovation. Die Zuverlässigkeit wird operationalisiert als Vertrauen in Ergebnisse sowie adäquate Qualität und Sicherheit während des gesamten Lebenszyklus von KI-Systemen. Zur sicheren Innovation wird ergänzt, dass, soweit angemessen, kontrollierte Umgebungen zur Entwicklung, Erprobung und Testung von KI-Systemen unter Behördenaufsicht ermöglicht werden sollen.

2.5. Kapitel IV – Rechtsmittel (Artikel 14–15)

[23] Vertragsparteien müssen unter Artikel 14 Massnahmen sicherstellen, um zugängliche und wirksame Rechtsmittel für Menschenrechtsverletzungen durch KI-Aktivitäten bereitzustellen, einschliesslich Dokumentation relevanter Informationen und Bereitstellung von Beschwerdemöglichkeiten. Artikel 15 fordert, dass bei erheblichen Menschenrechtsbeeinträchtigungen durch KI-Systeme wirksame Verfahrensgarantien verfügbar sind und Personen informiert werden, wenn sie mit KI-Systemen interagieren anstatt mit Menschen.

2.6. Kapitel V – Bewertung und Minderung von Risiken und nachteiligen Auswirkungen (Artikel 16)

[24] Vertragsparteien sollen gemäss den in Kapitel III dargelegten Prinzipien Massnahmen ergreifen zur Identifikation, Bewertung, Vermeidung und Minderung von Risiken durch KI-Systeme, mit einem Fokus auf tatsächliche und potenzielle Auswirkungen auf Menschenrechte, Demokratie und Rechtsstaatlichkeit. Diese Massnahmen sollen abgestuft und differenziert sein, und:

a. den Kontext und die beabsichtigte Nutzung der KI-Systeme berücksichtigen,
b. die Schwere und Wahrscheinlichkeit potenzieller Auswirkungen berücksichtigen,
c. die Perspektiven relevanter Interessengruppen einbeziehen,
d. iterativ im gesamten Lebenszyklus des KI-Systems angewendet werden,
e. die Überwachung von Risiken und negativen Auswirkungen beinhalten,
f. die Dokumentation von Risiken und Auswirkungen sowie den Risikomanagementansatz umfassen,
g. bei Bedarf Tests der KI-Systeme vor deren erstmaliger Nutzung und bei wesentlichen Änderungen einfordern.

[25] Im Weiteren werden Massnahmen zur Dokumentation negativer Auswirkungen von KI-Systemen auf Menschenrechte gefordert. Bei Unvereinbarkeit bestimmter KI-Anwendungen mit der Achtung der Menschenrechte, der Funktionsweise der Demokratie oder der Rechtsstaatlichkeit sollen Vertragsparteien zudem ein Moratorium, Verbot oder andere Massnahmen prüfen.

2.7. Kapitel VI – Implementierung der Konvention (Artikel 17–22)

[26] Bei der Implementierung der Konvention sollen Vertragsparteien besondere Sorge darauf legen, dass dies ohne Diskriminierung geschieht. Die speziellen Bedürfnisse von Menschen mit Behinderungen und Kindern sind zu berücksichtigen. Wichtige Fragen zu KI-Systemen sollen öffentlich diskutiert werden (inklusive Multistakeholder-Konsultationen). Die Förderung digitaler Bildung und Fähigkeiten für alle ist zu ermutigen. Bestehende Menschenrechte und rechtliche Verpflichtungen bleiben unberührt und die Vertragsparteien können weitergehenden Schutz gewähren, als in der Konvention festgelegt.

2.8. Kapitel VII – Follow-up-Mechanismus und Zusammenarbeit (Artikel 23–26)

[27] Die Konferenz der Vertragsparteien wird in Artikel 23 erläutert und besteht aus Vertretern der Vertragsparteien. Sie soll die effektive Anwendung und Umsetzung der Konvention fördern, Probleme identifizieren, Änderungen erwägen, Informationen austauschen und die Kooperation mit Stakeholdern ermöglichen, inklusive öffentlichen Anhörungen. Sie wird vom/von der Generalsekretär/in des Europarats einberufen und erstellt eigene Verfahrensregeln innert zwölf Monaten nach Inkrafttreten der Konvention. Die Vertragsparteien werden vom Europarat unterstützt, Expertise zur Unterstützung der Umsetzung der Konvention soll beigezogen werden können und Nicht-Mitglieder müssen zu den Aktivitäten finanziell beitragen.

[28] Unter Artikel 24 muss jede Vertragspartei innerhalb der ersten zwei Jahre nach Beitritt und danach regelmässig Berichte über ihre Massnahmen zur Umsetzung der Konvention vorlegen. Das Format und der Prozess der Berichterstattung werden von der Konferenz der Vertragsparteien festgelegt. Artikel 25 erwähnt die internationale Zusammenarbeit bei der Umsetzung der Konvention und hebt Risiken und Effekte hervor, die im Forschungskontext sowie in Bezug zur Privatwirtschaft aufkommen. Ein Aufruf zur Zusammenarbeit mit relevanten Interessengruppen wird ebenfalls erwähnt, um Risiken und nachteilige Auswirkungen auf Menschenrechte, Demokratie und Rechtsstaatlichkeit während des Lebenslaufs von KI-Systemen zu minimieren.

[29] Jede Vertragspartei muss unter Artikel 26 unabhängige und unparteiische Mechanismen zur Überwachung der Einhaltung der Konvention einrichten oder benennen. Diese Mechanismen sollen über die notwendigen Befugnisse, Fachkenntnisse und Ressourcen verfügen. Bei mehreren Mechanismen soll eine effektive Zusammenarbeit gefördert werden, auch mit bestehenden Menschenrechtsstrukturen.

2.9. Kapitel VIII – Schlussklauseln (Artikel 27–36)

[30] Im letzten Teil der Konvention werden noch zehn Aspekte abgehandelt. Auswirkungen der Konvention und Änderungen: Vertragsparteien können bestehende internationale Abkommen und Regelungen anwenden, solange sie mit den Zielen dieser Konvention übereinstimmen. Änderungen können von jeder Partei, dem Ministerkomitee des Europarats oder der Konferenz der Vertragsparteien vorgeschlagen werden. Änderungen treten in Kraft, nachdem alle Parteien zugestimmt haben.

[31] Die Konvention steht Mitgliedstaaten des Europarats, beteiligten Nichtmitgliedstaaten und der EU zur Unterzeichnung offen. Sie tritt drei Monate nach der Ratifizierung durch fünf Unterzeichner, darunter drei Europaratsmitglieder, in Kraft. Nach Inkrafttreten können Nichtmitgliedstaaten des Europarats auf Einladung des Ministerkomitees der Konvention beitreten.

3. Bevorstehende Herausforderungen

[32] Über 50 Staaten konnten sich auf eine Konvention einigen für einen Umgang mit KI, die Menschenrechte, Demokratie und Rechtsstaatlichkeit schützt – das ist ein wichtiges Signal mit globaler Strahlkraft.

[33] Es handelt sich bei diesem Text aber natürlich um einen Kompromiss. Den Vertragsparteien wird viel Spielraum bei der Interpretation und Umsetzung der Konvention gelassen; wie viel Wirkung sie tatsächlich entfalten kann, hängt massgeblich von den Deklarationen der Vertragsparteien ab sowie von der konkreten Durchsetzung auf Länderebene. Es folgen weitere Bemerkungen zu bevorstehenden Herausforderungen, in der Reihenfolge der Konventionsartikel.

3.1. Relevanter Geltungsbereich

[34] Der Spielraum für die Umsetzung der Konvention ist für Vertragsparteien in Artikel 3 am grössten. Mittels einer Deklaration können sie festlegen, inwiefern die Konvention auf die Privatwirtschaft angewendet wird. Das Hauptargument für diese Formulierung sind die unterschiedlichen Rechtssysteme, in denen die Konvention zum Einsatz kommen soll.

[35] Natürlich ist es sehr wichtig, dass Regierungen die Menschenrechte und Rechtsstaatlichkeit einhalten, wenn sie KI-Systeme einsetzen. Aber die mit Abstand grösste Dynamik in der KI-Entwicklung findet aktuell in der Privatwirtschaft statt, angeführt von den weltgrössten Technologiekonzernen. Nur wenn die KI-Konvention es schafft, diese Konzerne in die Verantwortung zu nehmen, kann sie auch ihre grösste Wirkung entfalten. Da die meisten dieser Konzerne ihren Hauptsitz in den USA haben, wäre die Unterzeichnung und Ratifizierung durch die USA, mit einer entsprechenden Deklaration zum Einbezug der Privatwirtschaft, ein grosser Erfolg. Auf eine Ratifizierung durch die USA kann man aktuell nur hoffen – und auch dann scheint ein Einbezug der Privatwirtschaft unwahrscheinlich.

[36] Weitere Spannung erzeugt Artikel 3 mit der Ausklammerung von nationalen Sicherheitsinteressen, hier ist einer schleichenden Ausweitung dieser Ausnahme früh ein Riegel vorzuschieben, was sich in der Umsetzung als Herausforderung gestalten dürfte. Die Ausnahme von Forschungs- und Entwicklungsaktivitäten steht ebenfalls in Spannung zu dem Einbezug des gesamten Lebenszyklus von KI-Systemen, auf den sich der Rest der Konvention jeweils bezieht. Dies ist von besonderer Relevanz, weil sich gewisse Aspekte von Risikominderung nur in der Entwicklungsphase sinnvoll umsetzen lassen.

3.2. Wirksame Prinzipien und Risikomanagement zum Umgang mit KI-Systemen

[37] Die Vertragsparteien verpflichten sich zur Förderung der Vertrauenswürdigkeit von KI-Systemen, indem sie die in Kapitel III dargelegten Massnahmen umsetzen. Die dazu erforderlichen technischen Lösungen stehen aber aktuell nicht bereit und deren Komplexität sollte nicht unterschätzt werden. Dies wird durch die Tatsache zusätzlich erschwert, dass sich der absolute Grossteil der Investitionen in KI auf neue und leistungsfähigere Modelle fokussiert. Gemäss einer Analyse von Prof. Markus Christen an der Universität Zürich zeigen Umfragen vor und nach der Einführung von ChatGPT, dass die Skepsis gegenüber der Nutzung von KI gestiegen ist.[5]

[38] Insbesondere bei KI-Systemen mit allgemeinem Verwendungszweck, wie z.B. Chat-GPT, ist es schwierig, Vertrauenswürdigkeit aufzubauen, da diese Systeme als Black-Box-Modelle fungie-

[5] Markus Christen, DSI Insights: ChatGPT erhöht die Skepsis gegenüber KI – darauf verzichten will man aber doch nicht (https://www.dsi.uzh.ch/de/current/news/2024/dsi-insights-ki-schweiz-skepsis.html).

ren, deren interne Entscheidungsprozesse schwer nachvollziehbar sind. Dies führt zu Bedenken hinsichtlich der Transparenz und Nachvollziehbarkeit der Ergebnisse. Zudem besteht das Risiko, dass diese Systeme unbeabsichtigt voreingenommene oder diskriminierende Antworten liefern, da sie auf grossen Datenmengen trainiert werden, die bestehende Vorurteile und Ungleichheiten widerspiegeln. Ein weiterer Faktor ist die Dynamik und Komplexität der Interaktionen, die solche KI-Systeme eingehen können, was die Kontrolle und Überwachung erschwert. Schliesslich gibt es Bedenken hinsichtlich des Datenschutzes und der Sicherheit, da KI-Systeme oft grosse Mengen an persönlichen Daten verarbeiten, was das Vertrauen der Nutzer:innen weiter untergräbt, wenn die Schutzmechanismen nicht ausreichend transparent und robust sind.

[39] Als Auswahl werden hier einige Massnahmen erwähnt, die in den Kapiteln III, IV und V eingefordert werden, zu denen es aber für die Umsetzung noch grössere technische Hürden gibt:

- Identifizierung von durch KI generierten Inhalten: Verschiedene Watermarking-Methoden wurden bisher getestet, es gibt aktuell keine, die ausreichend robust ist.[6]
- Vertragsparteien müssen Massnahmen ergreifen, um Verantwortlichkeit und Haftung für negative Auswirkungen auf Menschenrechte, Demokratie und Rechtsstaatlichkeit durch KI-Aktivitäten sicherzustellen. Der Weg zur Umsetzung dieser juristischen Herausforderung scheint gangbar, wenn auch noch nicht beschritten.
- Diskriminierungsverbot: Mit der aktuellen Modellarchitektur von KI-Systemen mit allgemeinem Verwendungszweck bisher nicht robust verhinderbar.[7]
- Zuverlässigkeit und Sicherheit: Auch die neusten KI-Systeme mit allgemeinem Verwendungszweck begehen erstaunliche Fehler und sind für Missbrauch anfällig.[8] Die technischen Herausforderungen für Erklärbarkeit und für robuste Evaluierung dieser Systeme sind gross, die Forschung dazu steckt in den Kinderschuhen.[9]
- Sichere Innovation: Ein aus Risikomanagement-Perspektive erstrebenswerter Ansatz ist «Safety by Design», also dass die Sicherheit von KI-Systemen auch in die ersten Forschungs- und Entwicklungsaktivitäten einbezogen wird. Dies ist bei der Architektur aktueller KI-Systeme mit allgemeinem Verwendungszweck nicht gegeben.[10] Entwickler:innen müssen nachweisen können, dass ihre KI-Systeme sicher sind.[11]

[40] Um die erwähnten Massnahmen auch umsetzen zu können, braucht es also signifikante Investitionen in die Sicherheit und Transparenz von KI-Systemen. Dabei kann nicht davon ausgegangen werden, dass diese Investitionen in ausreichendem Masse von den Entwicklerfirmen

---

6 Nikola Jovanović/Robin Staab/Martin Vechev, Watermark Stealing in Large Language Models (https://watermark-stealing.org/).

7 Valentin Hofmann/Pratyusha Ria Kalluri/Dan Jurafsky Sharese King, Dialect prejudice predicts AI decisions about people's character, employability, and criminality (https://arxiv.org/pdf/2403.00742).

8 Prabhakar Raghavan, Gemini image generation got it wrong. We'll do better. (https://blog.google/products/gemini/gemini-image-generation-issue/); Andy Zou et al, Universal and Transferable Adversarial Attacks on Aligned Language Models (https://llm-attacks.org/).

9 Usman Anwar et al., Foundational Challenges in Assuring Alignment and Safety of Large Language Models (https://arxiv.org/pdf/2404.09932).

10 Thomas Woodside/Helen Toner, How Developers Steer Language Model Outputs: Large Language Models Explained, Part 2 (https://cset.georgetown.edu/article/how-developers-steer-language-model-outputs-large-language-models-explained-part-2/).

11 David Dalrymple et al., Towards Guaranteed Safe AI:A Framework for Ensuring Robust and Reliable AI Systems (https://arxiv.org/pdf/2405.06624).

selbst getätigt werden und die Staaten selbst müssen hier Forschungsprojekte anstossen. Ob das Impact Assessment Werkzeug, das der Europarat in Begleitung der Konvention bis Ende 2024 erarbeitet, diesen Herausforderungen gewachsen ist, wird sich noch zeigen. Unabhängig davon wären Vertragsparteien gut beraten, die KI-Forschung und -Entwicklung genau zu verfolgen, in das Ökosystem für externe Sicherheitsbewertungen zu investieren und externe Audits von KI-Modellen mit allgemeinem Verwendungszweck vor der Inbetriebnahme dieser Modelle in Auftrag zu geben, wie es in anderen Branchen mit ausgereiftem Risikomanagement Standard ist.[12]

3.3. Follow-up Mechanismus und internationale Kooperation

[41] Die Umsetzung und Einhaltung der Konvention steht vor mehreren Herausforderungen. Erstens erschweren die unterschiedlichen rechtlichen und regulatorischen Rahmenbedingungen in den Vertragsstaaten die einheitliche Umsetzung der Konvention. Es ist daher notwendig, dass alle Vertragsparteien über die notwendigen Mittel und Kapazitäten verfügen, um ihre Verpflichtungen zu erfüllen und effektive Aufsichtsmechanismen zu etablieren. Ein weiterer wichtiger Aspekt ist die Sicherstellung, dass öffentliche Konsultationen und Dialoge mit verschiedenen Interessengruppen stattfinden und effektiv sind, um die vielfältigen Perspektiven und Bedenken angemessen zu berücksichtigen. Solche Massnahmen sind wichtig, um zu verhindern, dass die Konvention zu einem Papiertiger wird.

[42] Die internationale Zusammenarbeit spielt ebenfalls eine wichtige Rolle bei der Umsetzung der Konvention, stellt aber auch eine erhebliche Herausforderung dar, insbesondere wenn es darum geht, Nicht-Mitgliedstaaten einzubinden, allen voran China, das technologisch sowie geopolitisch ein zentraler Akteur ist. Eine effektive Koordination zwischen den Vertragsparteien ist unerlässlich, um eine harmonisierte Anwendung der Konvention zu gewährleisten. Dazu gehört auch der effektive Austausch relevanter Informationen und bewährter Verfahren zwischen den Staaten, beispielsweise durch KI-Vorfallmeldungen.

[43] Schliesslich sind Sanktionen und die Durchsetzung der Bestimmungen wesentliche Elemente, um die Einhaltung der Konvention sicherzustellen. Es müssen wirksame Mechanismen zur Sanktionierung und Durchsetzung bei Nichteinhaltung entwickelt werden.

[44] Die Bewältigung dieser Herausforderungen erfordert kontinuierliche Anstrengungen, Zusammenarbeit und Engagement seitens aller beteiligten Akteure. Die Schweiz kann mit gutem Vorbild vorangehen und den Worten Taten folgen lassen.

David Marti leitet bei dem Schweizer Think Tank *Pour Demain* das KI-Programm. Im Dialog mit Wissenschaft, Politik, Zivilgesellschaft und Wirtschaft entwickelt sein Team Empfehlungen für den Umgang mit KI auf nationaler und internationaler Ebene. Der Autor dankt Jacob Schaal und Patrick Stadler für die Mitarbeit an diesem Text.

12 Jess Whittlestone/Jack Clark, Why and How Governments Should Monitor AI Development (https://arxiv.org/pdf/2108.12427); Toby Shevlane et at, Model evaluation for extreme risks (https://arxiv.org/pdf/2305.15324); Merlin Stein and Connor Dunlop, Safe before sale (https://www.adalovelaceinstitute.org/report/safe-before-sale/).