eCH-0272 – Transparenz, Erklärbarkeit und Risiken von KI-Systemen

Name	Transparenz, Erklärbarkeit und Risiken von KI-Systemen
eCH-Nummer	eCH-0272
Kategorie	Standard
Reifegrad	Definiert
Version	1.0.0
Status	Entwurf
Beschluss am	
Ausgabedatum	2024-06-03
Ersetzt Version	Neu
Voraussetzungen	
Beilagen	
Sprachen	Deutsch (Original), Französisch (Übersetzung)
Autoren	Fachgruppenleiter: Robin Pekerman Mitwirkende: Anna Mätzener, Heidi Ates, Mark Strauch, Mevlüt Polat, Robin Pekerman, Ursulina Kölbener sowie in Zusammenarbeit mit BAKOM, BFS
Herausgeber / Vertrieb	Verein eCH, Räffelstrasse 20, 8045 Zürich T 044 388 74 64, F 044 388 71 80 www.ech.ch / info@ech.ch

Zusammenfassung

Bei dem vorliegenden Standard handelt es sich um einen konzeptionellen Standard, der Mindestanforderungen an die Transparenz, Erklärbarkeit und Risiken der KI-Systeme definiert. Aufgrund der umfassenden Analyse der Rahmenbedingungen, erarbeiteten Methoden und Ergebnisse gilt dieser Standard als Basisstandard für die weiteren Standards der Fachgruppe KI eCH.

Inhaltsverzeichnis

Hinweis

Im vorliegenden Dokument wird bei der Bezeichnung von Personen eine geschlechtsneutrale Formulierung verwendet. Basis bildet der Leitfaden der Bundeskanzlei. Je nach Situation kommen Paarformen (Bürgerinnen und Bürger), geschlechtsabstrakte Formen (versicherte Person), geschlechtsneutrale Formen (Versicherte) oder Umschreibungen ohne Personenbezug zum Einsatz. Das generische Maskulin (Bürger) ist nicht zulässig. Vollformen werden in fortlaufenden Texten verwendet, also in Texten, die aus ausformulierten Sätzen bestehen. In verknappten Textpassagen, namentlich in Tabellen, können Kurzformen verwendet werden. Dabei wird die Kurzform mit Schrägstrich, aber ohne Auslassungsstrich verwendet (Referent/in). Genderstern und ähnliche Schreibweisen werden nicht verwendet.

1 Einleitung

1.1 Status

Entwurf: Das Dokument wurde von den zuständigen Referenten aus dem Expertenausschuss zur öffentlichen Konsultation freigegeben und entsprechend publiziert.

1.2 Anwendungsgebiet

Der vorliegende Standard spezifiziert die Mindestanforderungen an Transparenz, Erklärbarkeit und Risiken von KI-Systemen und deren Auswirkungen. Der Standard orientiert sich an den Standards des Verein eCH in der Schweiz und definiert Vorgaben für die Entwicklung, Anwendung als auch für die Dokumentation von KI-Systemen unter Berücksichtigung der normativen Referenzen der schweizerischen Bundesverwaltung (insbesondere Leitlinien «Künstliche Intelligenz» für den Bund) sowie an weitere im Einzelfall anwendbare Gesetze (z.B. europäischen KI-Verordnung, EU AI-Act).

Der Nutzen dieses Standards liegt in der strukturierten Vorgabe von Mindestanforderungen, die sicherstellen sollen, dass KI-Systeme durch die Einhaltung von anwendbaren Gesetzen (z.B. Datenschutzgesetz und Urheberrechtsgesetz vgl. Kapitel 2) vertrauenswürdig gestaltet sind sowie ethische Prinzipien gewährleisten. Zudem unterstützt das Dokument öffentliche und private Institutionen bei der Implementierung und Überwachung von KI-Systemen, was zur Stärkung der Innovationskraft und Wettbewerbsfähigkeit der Schweiz beiträgt.

Beim vorliegenden Standard wurde nebst einer soliden Analyse der bestehenden gesetzlichen Rahmenbedingungen auch internationale Standards geprüft (vgl. Kapitel 6). Die Unterschiede von eCH-0272 zu den internationalen Standards sind wie folgt:

- Schweizer Spezifika: Der vorliegende Standard berücksichtigt spezifische schweizerische Gesetze und Rahmenbedingungen, die in internationalen Standards nicht im Detail behandelt werden.
- Praktische Anwendung: Der Fokus liegt auf der praktischen Anwendbarkeit in der Schweiz, einschliesslich spezifischer Empfehlungen für öffentliche und private Institutionen.
- ISO-Standards: ISO-Standards wie ISO/IEC 22989 und ISO/IEC 23053 fokussieren sich auf die Anforderungen an Transparenz und Erklärbarkeit, bieten jedoch eine allgemeine Anleitung für eine Vielzahl von Anwendungen und Sektoren. Diese Standards verfolgen einen allgemeinen risikobasierten Ansatz zur Identifizierung und Bewertung von Risiken bei der Implementierung von KI-Systemen.
- eCH-0272: Betont die Transparenz und Erklärbarkeit von KI-Systemen mit spezifischen Anforderungen und Empfehlungen, die an die Bedürfnisse und gesetzlichen Vorgaben der Schweiz angepasst sind. eCH-0272 orientiert sich ebenfalls an einem risikobasierten Ansatz, legt jedoch spezifische Massnahmen für die schweizerische Gesetzgebung fest, einschliesslich des Datenschutzes und der Wahrung der Grundrechte.

eCH-0272 soll als konzeptioneller Basisstandard eingesetzt werden, der eine solide Grundlage für die weiteren KI-Standards von eCH schafft. Er bietet eine solide Übersicht über das Umfeld der KI-Normierung und zeigt, wie die Rückverfolgbarkeit (Traceability) der Ergebnisse sichergestellt werden kann.

1.3 Geltungsbereich des Standards

Der vorliegende Standard richtet sich primär an Bund, Kantone und Gemeinden. Öffentlich-rechtliche Anstalten sowie Unternehmen aus der Privatwirtschaft können diesen Standard fakultativ ebenfalls anwenden.

Die genaueren Zielgruppen sind wie folgt beschrieben;

Dieser Standard richtet sich an alle Stakeholder, die in Kapitel 5 ausführlich beschrieben werden. Insbesondere umfasst dies die folgenden Gruppen:

Anbietende: Natürliche oder juristische Personen, Behörden, Einrichtungen oder andere Stellen, die ein KI-System entwickeln oder entwickeln lassen, um es unter ihrem eigenen Namen oder ihrer Marke, entweder entgeltlich oder unentgeltlich, auf den Markt zu bringen oder in Betrieb zu nehmen.

Nutzende: Natürliche oder juristische Personen, Behörden, Einrichtungen oder andere Stellen, die ein KI-System eigenverantwortlich verwenden. Dies gilt nicht, wenn das KI-System ausschliesslich im Rahmen einer persönlichen und nicht beruflichen Tätigkeit genutzt wird.

Regulierungsbehörden: Behörden, die KI-Systeme beaufsichtigen oder die Gesetzeskonformität der KI-Systeme prüfen.

1.4 Notation des Standards

Die Anforderungen in den Standards werden gemäss der Terminologie aus [RFC2119, Standard] spezifiziert, dabei kommen die folgenden Ausdrücke zur Anwendung, die durch GROSSSCHREIBUNG als Wörter mit den folgenden Bedeutungen kenntlich gemacht werden:

ZWINGEND: Die verantwortliche Person muss die Vorgabe umsetzen.

EMPFOHLEN: Die verantwortliche Person kann aus wichtigen Gründen auf eine Umsetzung der Vorgabe verzichten.

OPTIONAL: Es ist der verantwortlichen Person überlassen, ob sie die Vorgabe umsetzen will.

1.5 Lebenszyklus des Standards

Änderungen für diesen Standard können jederzeit mittels eines Änderungsantrags gemäss eCH003 erfolgen, welche die Bestimmungen für Minor und Major Changes definiert. Ausserdem soll dieser Standard alle fünf Jahre als Ganzes geprüft werden, ob dieser Standard noch im Einsatz ist und den aktuellen Bedürfnissen der Anwender entspricht. Weitere Bestimmungen betreffend dem Lebenszyklus sind aus den Standards eCH-0150 (Change und Release Management von eCH-Standards) und eCH-0218 (Life-Cycle-Management Fachgruppen) zu entnehmen.

2 Gesetzliche Grundlagen

Zum Zeitpunkt der Entstehung dieses Standards gibt es in der Schweiz keine spezifischen Gesetze über die Künstliche Intelligenz (KI). Allerdings gibt es Gesetze und Verordnungen (z.B. das Bundesgesetz über den Datenschutz), deren Geltungsbereich auch KI umfasst.

Der Bundesrat hat sich mit den Entwicklungen, Chancen und Herausforderungen von KI befasst. Er hat das Eidgenössische Departement für Umwelt, Verkehr, Energie und Kommunikation (UVEK) beauftragt, bis Ende 2024 mögliche Ansätze zur Regulierung von KI aufzuzeigen und dabei alle Bundesstellen miteinzubeziehen, die bei den betroffenen Rechtsbereichen federführend sind.

In der Europäischen Union (nachfolgend EU) hatte die EU-Kommission im Rahmen der EU-Digitalstrategie mit dem Artificial Intelligence Act (AIA) einen Entwurf über ein Gesetz über Künstliche Intelligenz publiziert. Der Entwurf enthält in dieser Form konkrete Vorschläge zur Regelung im Umgang mit Künstlicher Intelligenz (KI) in der Forschung und Wirtschaft.

Die nachfolgend aufgeführte Auflistung der gesetzlichen Grundlagen in der Schweiz und der EU ist nicht abschliessend.

2.1 Gesetze und Verordnungen in der Schweiz

Bundesverfassung der Schweizerischen Eidgenossenschaft (BV, SR 101)

Die Grundrechte sind in den Artikeln 7 ff. der Bundesverfassung (wie z.B. die Menschenwürde, der Schutz der Privatsphäre und der Persönlichkeitsschutz) aufgeführt und müssen in der gesamten Rechtsordnung eingehalten werden. Beim Einsatz von KI kommt dem Schutz der Grundrechte besondere Bedeutung zu. Grundrechtliche und ethische Überlegungen müssen bei der Gestaltung und der Anwendung von KI beachtet werden («ethics by design»).

Bundesgesetz über den Datenschutz (Datenschutzgesetz, DSG; SR 235.1)

Dieses Gesetz bezweckt den Schutz der Persönlichkeit und der Grundrechte von natürlichen Personen, über die Personendaten bearbeitet werden. Das DSG ist auf den Einsatz von KI-gestützten Datenbearbeitungen direkt anwendbar1. Hersteller, Anbieter und Verwender entsprechender KI-Systeme müssen sich bei der Bearbeitung von Personendaten an das DSG halten wie z.B. die Informationspflicht bei einer automatisierten Einzelentscheidung oder die Durchführung einer Datenschutz-Folgeabschätzung bei einer Datenbearbeitung, der ein hohes Risiko für die Persönlichkeitsrechte oder Grundrechte der betroffenen Personen mit sich bringt. Das hohe Risiko ergibt sich, insbesondere bei Verwendung neuer Technologien, aus der Art, dem Umfang, den Umständen und dem Zweck der Bearbeitung.

[1] Mitteilung Eidgenössischer Datenschutzbeauftragter vom 09. November 2023.

Bundesgesetz über das Urheberrecht und verwandte Schutzrechte (Urheberrechtsgesetz, URG; SR 231.1)

Das Urheberrechtsgesetz schützt das geistige Eigentum eines Urhebers. Für KI ist der Bereich des geistigen Eigentums bei der Bearbeitung oder Nutzung von Daten, die durch das Urheberrecht geschützt sind, ebenfalls wichtig. Wenn mit KI-Systeme Inhalte erstellt oder verwendet werden, die urheberrechtlich geschützt sind, kommt dieses Gesetz zur Anwendung.

Schweizerische Strafgesetzbuch (StGB; SR 311.0) und weitere anwendbare Gesetze

Entstehen durch KI-Schäden, so haftet eine haftet eine natürliche oder juristische Person für durch KI entstandene Schäden, da Roboter keine Rechtspersönlichkeit haben, wenn die Haftungsvoraussetzungen erfüllt sind. Es gilt in solchen Fällen das StGB sowie unter anderem das Obligationenrecht (OR; SR 220) oder das Bundesgesetz über die Produktehaftpflicht (Produktehaftpflichtgesetz, PrHG; SR 221.112.944).

2.2 Gesetze und Verordnungen in der EU und Europarat

Im Rahmen der Entwicklung dieses Standards wurden relevante internationale Gesetze und Abkommen eingehend analysiert und geprüft. Es wurde darauf geachtet, dass einzelne Aspekte dieser Regelwerke in den verschiedenen Themenbereichen des Standards Berücksichtigung finden. Es ist jedoch zu beachten, dass ausschliesslich jener Bestimmungen aus diesen Gesetzen und Abkommen in den Standard aufgenommen werden, die mit den schweizerischen Normen übereinstimmen und den spezifischen Anforderungen und Bedürfnissen dieses Standards entsprechen können. Jegliche Übernahme von Regelungen erfolgt somit unter dem Vorbehalt der Kompatibilität und Relevanz für die schweizerischen Rahmenbedingungen. Die vollständige und vertiefte Behandlung dieser internationalen Regelwerke konnte für zukünftige Versionen des Standards vorgesehen.

2.2.1 KI-Verordnung der EU (AI Act) zur Regulierung der künstlichen Intelligenz

Die EU-KI-Verordnung stellt einen umfassenden Rechtsrahmen auf, der klare Regeln und Standards für den Umgang mit Künstlicher Intelligenz festlegt, um die Grundrechte zu schützen. Diese Regelungen gewährleisten, dass KI-Systeme die Würde, Privatsphäre und andere fundamentale Rechte der Einzelpersonen respektieren und nicht, diskriminierend, manipulativ oder anderweitig schädlich agieren.

Zur Stärkung des Vertrauens in KI-Technologien fordert die Verordnung, dass KI-Systeme nachvollziehbar und erklärbar sind. Sie beinhaltet daher Bestimmungen zur Transparenz, Rechenschaftspflicht und Verständlichkeit.

In der Verordnung wird eine Einteilung in vier Risikoklassen vorgenommen: niedriges, begrenztes, hohes und inakzeptables Risiko. KI-Systeme, die ein inakzeptables Risiko darstellen, sind verboten. Die rechtlichen Anforderungen variieren je nach Risikoklasse, wobei für Hochrisiko-Systeme besonders strenge Vorschriften gelten. Entwickler solcher Systeme sind verpflichtet, eine umfassende Risikobewertung durchzuführen.

Zudem sind Dokumentationspflichten und die Einhaltung bestimmter technischer Standards vorgesehen. Der Zweck dieser Verordnung ist die Verbesserung des Funktionierens des Binnenmarkts durch die Schaffung eines einheitlichen Rechtsrahmens, der insbesondere die Entwicklung, Vermarktung und Nutzung von Künstlicher Intelligenz im Einklang mit den Werten der EU regelt.

Alle Akteure, die unter den AI Act fallen, sollen Hochrisiko-KI-Systeme im Einklang mit den folgenden sechs «KI-Grundsätzen» entwickeln und einsetzen (Europäische Kommission, 2023):

- Menschliches Handeln und Kontrolle: KI-Systeme sollen dem Menschen dienen und die Menschenwürde sowie persönliche Autonomie respektieren, und so funktionieren, dass sie von Menschen kontrolliert und überwacht werden können.
- Technische Robustheit und Sicherheit: Unbeabsichtigte und unerwartete Schäden sollen auf ein Mindestmass reduziert werden und KI-Systeme sollen im Falle von unbeabsichtigten Problemen robust sein.
- Data Gouvernance: KI-Systeme mit hohem Risiko, die Techniken verwenden, bei denen Modelle mithilfe von Daten trainiert werden, müssen die Trainings-, Validierungs- und Testdatensätze relevant, repräsentativ, fehlerfrei und vollständig sein.
- Transparenz: Es muss eine Rückverfolgbarkeit und Erklärbarkeit möglich sein und den Menschen muss bewusst gemacht werden, dass sie mit einem KI-System interagieren.
- Vielfalt, Nichtdiskriminierung und Fairness: KI-Systeme sollen unterschiedliche Akteure einbeziehen und den gleichberechtigten Zugang, die Gleichstellung der Geschlechter und die kulturelle Vielfalt fördern, und umgekehrt diskriminierende Auswirkungen vermeiden.
- Soziales und ökologisches Wohlergehen: KI-Systeme sollen nachhaltig und umweltfreundlich sein sowie zum Nutzen aller Menschen entwickelt und eingesetzt werden.

2.2.2 KI Konvention Europarat

Die KI-Konvention des Europarates hat den Auftrag, das erste verbindliche internationale Abkommen zu KI zu verhandeln, welches auf den Normen des Europarats zu Menschenrechten, Demokratie und Rechtsstaatlichkeit beruht. Dieses Abkommen soll nicht nur die Einhaltung dieser Grundwerte sicherstellen, sondern auch die Innovation in der Entwicklung und Anwendung von KI-Technologien fördern.

Anlässlich des 75-jährigen Bestehens des Europarats hat sein Ministerkomitee einstimmig den Text verabschiedet, der im März bereits vom Ausschuss für künstliche Intelligenz (Committee on Artificial Intelligence, CAI) angenommen worden war. Das Übereinkommen über KI wird im September 2024 für alle Staaten zur Unterzeichnung aufgelegt. Bei einer Ratifikation durch die Schweiz muss es in das innerstaatliche Recht überführt werden.

Das Übereinkommen schafft einen gemeinsamen und rechtsverbindlichen Rahmen für KI-Systeme. Dieser stellt unter anderem sicher, dass die Normen des Europarats und andere internationale Standards in Bezug auf Menschenrechte, Demokratie und Rechtsstaatlichkeit bei der Entwicklung und Nutzung dieser Systeme eingehalten werden. Dabei stützt es sich auf allgemeine Regeln und Grundsätze wie Transparenz, Robustheit, Nichtdiskriminierung und Schutz der Privatsphäre. Ausserdem fördert es über die Grenzen Europas hinaus Werte und einen gemeinsamen Rahmen, die es zu beachten gilt.

Das Übereinkommen nennt zunächst die Grundsätze für Tätigkeiten innerhalb des Lebenszyklus von KI-Systemen. Weiter definiert es die Einsprachemöglichkeiten und den Rahmen für die Beurteilung und Milderung von Risiken und nachteiligen Folgen. Die Umsetzung des Übereinkommens wird ebenfalls im Detail festgelegt und durch einen Monitoring- und Zusammenarbeitsmechanismus sowie durch Schlussbestimmungen ergänzt.

3 Terminologie

Die in diesem Dokument beschriebene Terminologie umfasst die am häufigsten benutzten Definitionen im Bereich der Künstlichen Intelligenz. Diese Begrifflichkeiten basieren auf den Arbeiten des beim Bundesamt für Statistik angesiedelten Kompetenznetzwerks für KI (CNAI). Für eine umfassende und detaillierte Darstellung der Terminologie verweisen wir auf die Webseite des Kompetenznetzwerks für KI (CNAI) sowie auf den Verhaltenskodex des Bundes für menschenzentrierte und vertrauenswürdige Datenwissenschaft, wo spezifische Definitionen und weiterführende Erläuterungen zu finden sind. Diese Terminologie ist nicht abschliessend und beim Einsatz von KI-Systemen können auch andere als die in Ziffer 3 aufgeführten Definitionen verwendet werden.

KI («Artificial Intelligence – AI»), heute manchmal als «maschinelle Intelligenz» («Machine Intelligence») bezeichnet, wird definiert als «einen Computer so bauen oder programmieren, um Dinge zu tun, die normalerweise menschliche oder biologische Fähigkeiten («Intelligenz») erfordern», z. B. visuelle Wahrnehmung (Bilderkennung), Spracherkennung, Sprachübersetzung, visuelle Übersetzung und Spiele spielen (mit konkreten Regeln).

KI-System: Ein KI-System ist ein maschinenbasiertes System, das für explizite oder implizite Ziele aus den empfangenen Inputs schlussfolgert, wie es Outputs wie Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erzeugen kann, welche die physische oder virtuelle Umgebung beeinflussen können. KI-Systeme können mit unterschiedlichem Ausmass an Autonomie ausgestattet werden.

Deep Learning: Eine Art von maschinellem Lernen basierend auf künstlichen neuronalen Netzwerken (Abstraktion von Informationsverarbeitung mittels künstlicher Neuronen), bei dem mithilfe eines mehrstufigen Verfahrens stufenweise detailliertere Eigenschaften von Daten extrahiert werden. (siehe Algorithmen im schweizerischen öffentlichen Diskurs: Eine Studie im Auftrag von AlgorithmWatch Schweiz)

Maschinelles Lernen (ML): Maschinelles Lernen (ML – «Machine Learning») ist ein weiteres Teilgebiet der KI, welches «den Computern die Lernfähigkeit verleiht». ML untersucht die Konstruktion von Algorithmen, die durch den Einsatz von Computern Daten analysieren und dabei automatisch lernen, sich anpassen und verbessern (anhand von konkreten vom Menschen vorgegebenen Regeln). Das resultierende statistische Modell ermöglicht bspw. Vorhersagen und Klassifizierungen von (noch nicht gesichteten) Daten, welche entscheidungsunterstützend eingesetzt werden können.

4 Lebenszyklusphasen und Stakeholder-Rollen KI-Systeme

Das KI-Systemlebenszyklusmodell in ISO/IEC22989 beschreibt die Entwicklung eines KI-Systems von der Initialisierung bis zur Ausserbetriebnahme. Das Modell besteht aus den folgenden sieben Lebenszyklusphasen: Initialisierung, Design und Entwicklung, Verifizierung und Validierung, Implementierung, Betrieb und Überwachung, Re-Evaluation und Ausserbetriebnahme (DIN SPEC 92001-3:2023-08).

Spezifische Prozesse und Zeitpläne können während einer oder mehrerer Lebenszyklusphasen auftreten und einzelne Phasen können wiederholt werden. Das Lebenszyklusmodell sollte als Hilfsmittel und nicht als Rezept betrachtet werden; Basierend auf unterschiedlichen Entwicklungstechniken oder Anwendungsdomänen sind unterschiedliche Lebenszyklusmodelle möglich.

In Anlehnung an ISO/IEC22989 identifiziert dieses Dokument zwölf verschiedene Stakeholder-Rollen, unterteilt in sechs Kategorien KI-Anbieter, KI-Hersteller, KI-Kunde, KI-Partner, KI-Subjekt und relevante Behörden. Eine einzelne Einheit oder Organisation kann gleichzeitig mehrere Stakeholder-Rollen übernehmen und eine einzelne Stakeholder-Rolle kann auf mehrere Organisationen verteilt sein.

Stakeholder auf verschiedenen Stufen werden möglicherweise auf eine von vier charakteristischen Arten mit KI-Systemen interagieren: entwickeln, bereitstellen, verwenden oder bewerten. Jede charakteristische Art der Interaktion bringt spezifische Informationsbedürfnisse mit sich.

	Stakeholder-Rollen											
Lebenszyklusphasen	**KI-Anbieter**		**KI-Hersteller**	**KI-Kunde**	**KI-Partner**				**KI-Subjekt**		**relevante Behörde**	
	KI-Plattform-Anbieter	KI-Produkt/Dienstleistung-Anbieter	KI-Entwickler	KI-Nutzer	KI-System Integrator	Daten Anbieter	KI-Bewerter	KI-Auditor	KI-Daten Subjekt	Andere Subjekte	Policy Erstel-ler	Regulator
Initialisierung	-	entwickeln	entwickeln	-	entwickeln	-	-	-	-	-	-	-
Design & Entwicklung	entwickeln	entwickeln	entwickeln	-	entwickeln	entwi-ckeln	-	-	-	-	-	-
Verifizierung & Validierung	-	entwickeln, bewerten	entwickeln, bewerten	-	entwi-ckeln, bewerten	-	bewerten	bewerten	-	-	bewer-ten	bewerten
Implementierung	entwickeln	bereitstellen, bewerten	entwickeln, bewerten	bereitstellen, verwenden	bereitstel-len, bewerten	-	-	-	-	-	-	-
Betrieb & Überwachung	bereitstellen	bereitstellen, bewerten	entwickeln, bewerten	bereitstellen, verwenden	bereitstel-len, bewerten	-	bewerten	bewerten	verwen-den	verwenden	-	-
Re-Evaluation	-	bewerten	-	-	-	-	bewerten	bewerten	-	-	bewer-ten	bewerten
Ausserbetriebnahme	-	entwickeln, bewerten	entwickeln, bewerten	-	-	-	-	-	-	-	-	-

Tabelle 1: Stakeholderrollen und Lebenszyklusphasen: Eigene Darstellung (übersetzt) in Anlehnung DIN SPEC 92001-3:2023-08, 2023)

5 Grundsätze für die KI in der Schweiz

5.1 «Leitlinien KI für die Bundesverwaltung»

Eine departementsübergreifende Arbeitsgruppe, die vom Staatsekretariat für Bildung, Forschung und Innovation (SBFI) gesteuert wurde, hat im Jahr 2019 einen Bericht zum Thema "Herausforderungen der Künstlichen Intelligenz" verfasst und auf dieser Basis die in Zusammenarbeit mit dem UVEK die Leitlinien «Künstliche Intelligenz» für die Bundesverwaltung erarbeitet.

In den Leitlinien kommen für den Umgang mit KI als Grundlage die für die Schweiz geltende nationale und internationale Rechtsordnung, insbesondere die Bundesverfassung der Schweizerischen Eidgenossenschaft (BV, SR 101) und die Normen der Europäischen Konvention zum Schutze der Menschenrechte und Grundfreiheiten (EMRK) und noch weitere gesetzliche Grundlagen zur Anwendung.

Diese Leitlinien bieten der Bundesverwaltung sowie den Trägern von Verwaltungsaufgaben des Bundes einen allgemeinen Orientierungsrahmen und sollen spezifisch in folgenden Kontexten beachtet werden:

- bei der Erarbeitung sektoraler KI-Strategien;
- bei der Einführung oder Anpassung von spezifischen, sektoralen Regulierungen;
- bei der Entwicklung und beim Einsatz von KI-Systemen innerhalb der Bundesverwaltung;
- bei der Mitgestaltung des internationalen Regelwerks zu KI.

Die Leitlinien zur künstlichen Intelligenz für den Bund werden angewendet und sind immer noch aktuell. Zu diesem Schluss kommt das BAKOM in seiner im Herbst 2022 durchgeführten Evaluation. Es sieht keinen Anpassungsbedarf. Die nächste Evaluation ist für 2024 geplant. Diese Leitlinien wurden im Rahmen des vorliegenden Standards konsultiert und bei den entsprechenden Themen berücksichtigt.

5.2 Verhaltenskodex des Bundes für die Datenwissenschaft und für vertrauenswürdige Datenräume

Eine weitere Referenz bildet der Verhaltenskodex des Bundes für die menschenzentrierte und vertrauenswürdige Datenwissenschaft. Die Grundprinzipien wie Daten- und Informationsschutz, Datensicherheit, Nicht-Diskriminierung, Nachvollziehbarkeit, Reproduzierbarkeit, Objektivität, Informationssicherheit, Daten-Gouvernanz, Erklärbarkeit, Transparenz, Neutralität, Ethischer Umgang mit Daten und Ergebnissen wurden im Rahmen des vorliegenden Standards konsultiert und bei den entsprechenden Themen berücksichtigt (BFS, 2023).

Der Verhaltenskodex für vertrauenswürdige Datenräume dient als Empfehlung oder Leitfaden für Akteure aus dem privaten Sektor, der Wissenschaft und der Zivilgesellschaft, die alle aufgefordert sind, ihn zu unterzeichnen. Der Kodex konkretisiert die Gestaltung vertrauenswürdiger Datenräume auf der Grundlage von vier Schlüsselprinzipien, nämlich Transparenz, Kontrolle, Fairness und Effektivität. Jeder dieser vier Grundsätze enthält eine Reihe praktischer Empfehlungen mit entsprechenden möglichen Umsetzungsmassnahmen. Die Gesellschaft als Ganzes würde von den Effizienzgewinnen und dem Innovationspotenzial profitieren, die der Datenaustausch mit sich bringt. Wir betrachten daher die digitale Selbstbestimmung als langfristiges Ziel.

6 Übersicht Internationale Standards und Normen

Bei der Erarbeitung der eCH KI-Standards wurden die internationalen Standards konsultiert. Aufgrund der eigenen Gesetzgebung und den landesspezifischen Bedürfnissen in der Schweiz differenzieren sich die eCH-Standards zu den internationalen Standards sowie es in den anderen Standards der eCH der Fall ist. Es ist von Bedeutung zu erwähnen, dass die Anforderungen, die in diesem Standard festgehalten sind, anhand internationaler Standards überprüft oder unterstützt werden können. Wenn erforderlich, werden inhaltliche Analogien im Standard berücksichtigt. Die Adressaten dieses Standards haben die Möglichkeit, zusätzlich zu den eCH-Standards auch internationale Standards anzuwenden, insbesondere wenn sie international tätig sind oder es für ihre Unternehmenstätigkeit erforderlich ist.

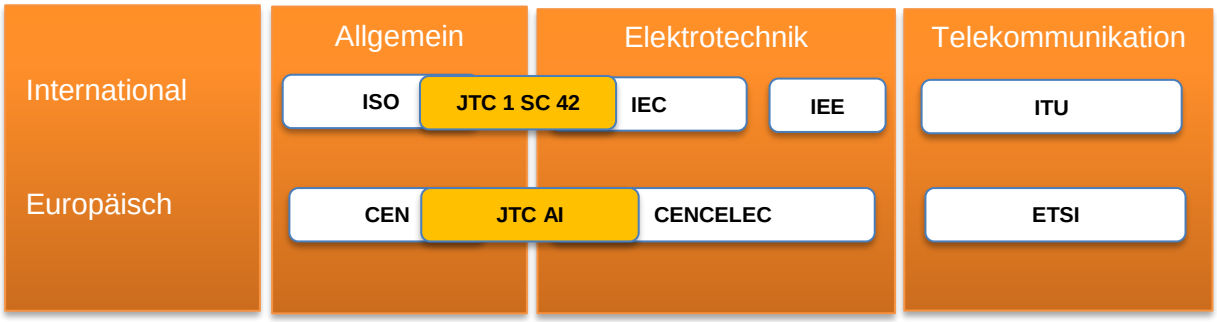

Tabelle 2: Übersicht internationale Standards: Eigene Darstellung

Folgend ist eine Auflistung von Standardisierungsarbeiten von ISO und IEC zu finden. Auf internationaler Ebene haben ISO und IEC ein gemeinsames Standardisierungskomitee für KI gegründet, nämlich ISO/IEC JTC 1/SC 42 (Subcommittee 42), das im Bereich KI und Big Data aktiv ist. Ihre Arbeitsgruppen (Working Groups) bereiten Standards zu verschiedenen Aspekten der KI vor, siehe folgende Tabelle.

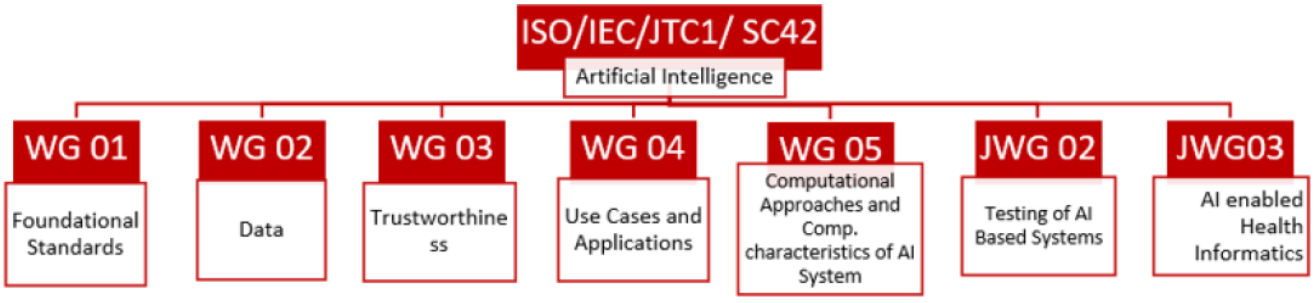

Abbildung 1: Übersicht ISO/IEC JTC 1/SC 42

Im Hinblick auf Standards in der KI spielen europäische Standards eine entscheidende Rolle bei der Unterstützung eines wirksamen KI-Regulierungsrahmensystems. In diesem Zusammenhang pflegen die Standardisierungsorganisationen einen engen und etablierten Dialog mit der Europäischen Kommission zu strategischen Fragen im Zusammenhang mit der KI-Standardisierung.

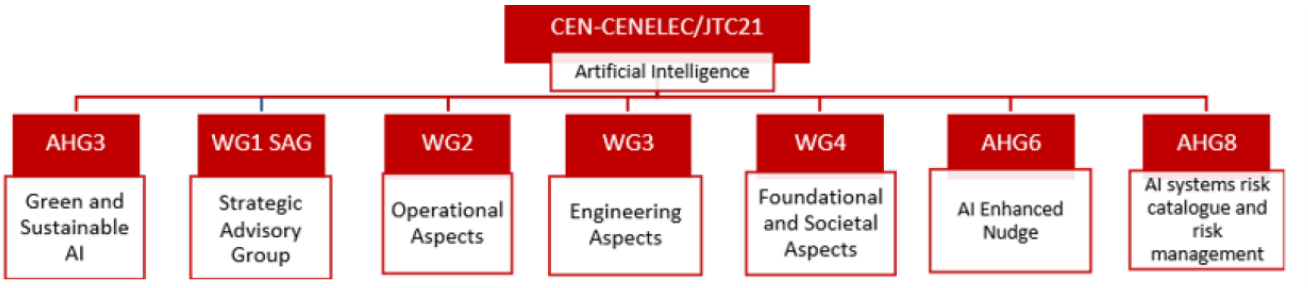

Abbildung 2: Übersicht CEN-CENELEC JTC 21

7 Anforderungen an KI-Systeme

Die nachfolgenden Anforderungen orientieren sich an einem risikobasierten Ansatz. Dieser Ansatz stützt sich inhaltlich auf die im Kapitel 2 "Gesetzliche Grundlagen", wo die rechtlichen Rahmenbedingungen detailliert dargelegt werden. Zudem werden im Kapitel 4 "KI Grundsätze Schweiz und EU" die Grundprinzipien für den Einsatz von Künstlicher Intelligenz beschrieben, welche bei der Erhebung von Anforderungen richtungsweisende Unterstützung leisten.

Folgend wird eine Übersicht dargestellt, zu welchen Risiken die nachfolgend beschriebenen Anforderungen und Massnahmen im Umgang mit künstlicher Intelligenz gestellt werden:

Risiko	**Analogien zu folgenden Normen und Regelwerke**
Grundrechte	Bundesverfassung, KI-Konvention, AI-Act
Datenschutz	Datenschutzgesetz
Urheberrechte	Urheberrechtgesetz
Autonomie	Leitlinien KI Bundesverwaltung, Verhaltenscodex Datenwissenschaft des Bundes
Fairness	Leitlinien KI Bundesverwaltung, Verhaltenscodex Datenwissenschaft des Bundes
Rechenschaftspflicht	Leitlinien KI Bundesverwaltung, Verhaltenscodex Datenwissenschaft des Bundes
Transparenz	Leitlinien KI Bundesverwaltung, Verhaltenscodex Datenwissenschaft des Bundes
Erklärbarkeit	Leitlinien KI Bundesverwaltung, Verhaltenscodex Datenwissenschaft des Bundes
Allgemein	Leitlinien KI Bund, Verhaltenscodex Datenwissenschaft des Bundes

Die Adressaten der einzelnen Anforderungen in den nächsten Kapiteln sind anhand des Attributs «KI-Lebenszyklus aufgezeigt. Für die Zuordnung der Adressdaten zu jedem KI-Lebenszyklus ist das Kapitel 5 «Life Cycle und Stakeholder-Rollen KI-Systeme» zu konsultieren.

Die in den nächsten Kapiteln beschriebenen Anforderungen und Massnahmen sind nicht abschliessend. Je nach Anwendungsfall und Kontext des KI-Systems können weitere Anforderungen bzw. Massnahmen ergriffen werden. Anhand der Notation wird aufgezeigt, ob eine Anforderung zwingend, empfohlen oder optional ist. Bei letzterem geht es um unterstützende Hilfsstellungen, falls keine andere Massnahme zur Problemstellung dient.

Ergänzende Informationen: Für die Erhebung von Anforderungen und Massnahmen an die genannten Risiken werden bei deren Beschreibungen folgende Parameter angewendet, um die Traceability (Rückverfolgbarkeit) sicherzustellen. Die Parameter werden exemplarisch mit der folgenden Tabelle angezeigt.

Beschreibung			**Traceability**		
			Explizit		Implizit
Nr.	**Notation**	**(Mindest)- Anforderung**	**Gesetz**	**Studie/ Bericht/Int. Standard**	**E-Input**
A1	ZWINGEND	Bei der Bearbeitung von Personendaten durch ein KI-System müssen die betroffenen Personen über den Zweck, über die Identität und Kontaktdaten des Verantwortlichen und gegebenenfalls über die Empfängerinnen und Empfänger der Personendaten informiert werden.	DSG	n/a	Ja/Nein?
A2	OPTIONAL	…….	n/a	n/a	Ja

Tabelle 3: Parameter (exemplarisch)

Die Nummerierung wird für die Identifikation der einzelnen Inhalte angezeigt. Die Notation bezieht sich auf die Vorgaben im Kapitel 1.4, welche üblicherweise in eCH-Standards angewendet wird. Die Anforderungen werden einzeln erfasst. Mit dem Traceability Parameter wird angezeigt, an welchen Quellen sich die Anforderungen anlehnen. Pro Anforderungen können eins oder mehrere Traceability Parameter gesetzt sein

7.1 Allgemein

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
A1	Design & Entwicklung, Verifikation & Validation	ZWINGEND	Der Anwendungsbereich des KI-Systems ist für die Bewertung der Risikoklassifikation erfasst werden.	Dokumentation des Anwendungsbereichs Erstellung einer detaillierten Dokumentation, die den beabsichtigten Anwendungsbereich des KI-Systems klar definiert. Dies sollten die Szenarien, in denen das System eingesetzt werden soll, die Zielbenutzergruppe und die spezifischen Aufgaben, die das KI-System ausführen wird, umfassen. Diese Informationen sollten öffentlich zugänglich gemacht werden.	Hoch
A2	Design & Entwicklung, Verifikation & Validation	ZWINGEND	Die Lernmethode (überwacht, unüberwacht, bestärkendes Lernen) der ML wird für die Bewertung der Risikoklassifikation erfasst. Der Grund liegt darin, dass bestimmte Lernmethoden, wie tiefe neuronale Netze (Deep Learning) sehr komplexe Modelle erzeugen können, die schwierig zu interpretieren und zu verstehen sind. Diese mangelnde Transparenz kann zu einer höheren Risikoklassifizierung führen, insbesondere wenn das KI-System in kritischen Anwendungsbereichen eingesetzt wird, in denen Entscheidungen nachvollziehbar sein müssen.	Implementierung von Audit-Protokollen Einrichtung umfassender Audit-Protokolle und Überwachungsprozesse, um sicherzustellen, dass das Verhalten des KI-Systems und seine Entscheidungsfindung kontinuierlich auf Übereinstimmung mit den vorgesehenen Funktionen und ethischen Richtlinien überprüft werden. Diese Protokolle können dazu beitragen, Bias, Fehlverhalten oder andere Risikofaktoren zu identifizieren und zu adressieren, was die Sicherheit und Vertrauenswürdigkeit des Systems erhöhten.	Hoch
A3	Design & Entwicklung, Verifikation & Validation,	OPTIONAL	Für die Bewertung der Risiken eines KI-Systems sollen messbare Werte erfasst werden. Mindestens sollten folgende Werte erfasst	Die Umsetzung dieser Anforderung soll in der Praxis mit vorhandenen Möglichkeiten	Tief

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
	Implementierung		werden: Wert für Risikoklassifikation sowie auch Wert für den Anwendungsbereich.	geprüft werden. Bei Bedarf kann folgendes Beispiel angewendet werden: Referenziertes Beispiel Anwendungsbereich (30%): Mindestanforderungen: a. Beschreibung des vorgesehenen Anwendungsbereichs. b. Identifikation möglicher negativer Auswirkungen in diesem Bereich. c. Vergleich mit bestehenden Lösungen (wenn vorhanden). Begründung: Der Anwendungsbereich kann erhebliche Auswirkungen auf die Art und Weise haben, wie ein KI-System Risiken verursachen kann. Ein KI-System, das in einem kritischen Bereich wie der Medizin eingesetzt wird, birgt potenziell grössere Risiken als eines, das in einem weniger kritischen Bereich eingesetzt wird. Lernmethode (20%): Mindestanforderungen: a. Beschreibung der Lernmethode und warum sie für das Problem geeignet ist. b. Analyse der Datenquellen und ihrer Qualität.	

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
				c. Bewertung von Bias- und Fairness-Fragen. Begründung: Die Lernmethode und die zugrunde liegenden Daten können zu versteckten Voreingenommenheit und Fehlern führen. Einige Methoden sind anfälliger für Überanpassung oder erfordern sorgfältigere Überprüfung als andere. Risikowert (50%): Mindestanforderungen: a. Quantifizierung des Gesamtrisikos basierend auf den beiden vorherigen Kategorien und anderen relevanten Faktoren. b. Beschreibung möglicher worst-case Szenarien. c. Notfallpläne und Massnahmen zur Risikominderung. Begründung: Das direkte Messen und Quantifizieren des Risikos geben einen klaren Hinweis auf die potenziellen Gefahren des Systems. Durch die Gewichtung mit 50% wird deutlich, dass das direkte Risikoprofil des KI-Systems von grösster Bedeutung ist. Die gesetzten prozentualen Zahlen repräsentieren die relative Wichtigkeit jeder Dimension. Der Anwendungsbereich ist wich-	

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
				tig, aber er wird durch den tatsächlichen Risikowert des Systems überlagert, daher nur 30%. Die Lernmethode ist zwar kritisch, aber weniger als die beiden anderen Kategorien, daher 20%. Der Risikowert hingegen, der direkte und umfassende Indikator für die Gefahr des Systems, hat den höchsten Prozentsatz von 50%.	
A4		OPTIONAL	Die gesellschaftlichen und ethischen Auswirkungen des Einsatzes eines KI-Systems sollen auch in die Bewertung seiner Risikoklassifikation einfliessen.	Einführung von Prozessen, um die gesellschaftlichen und ethischen Auswirkungen des Einsatzes von KI-Systemen festzustellen und zu adressieren	

Tabelle 4: Anforderung KI-Systeme – Allgemein

Ergänzende Informationen:

Die Traceability Parameter für die Anforderungen in diesem Kapitel sind wie folgt:

	Traceability		
Nr.	**Explizit**		**Implizit**
	Gesetz	***Studie/ Bericht/Int. Standard***	***Experten-Input***
A1	-	Leitlinien Künstliche Intelligenz Bund, AI-Act	Kernteam FG KI
A2	-	-	Kernteam FG KI
A3	-	-	Kernteam FG KI
A4	-	-	Kernteam FG KI

Tabelle 5: Anforderung KI-Systeme – Allgemein ergänzende Informationen

7.2 Grundrechte

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
A5	Design & Entwicklung, Betrieb & Überwachung, Re-Evaluation	ZWINGEND	Menschenzentrierung: Die Menschenwürde muss bei der Entwicklung und Anwendung von KI-Systemen geachtet werden. KI-Systeme sind mit dem primären Ziel zu entwickeln und einzusetzen, dass das Wohlergehen der Menschen gefördert wird. Sie sollen sich an den aktuellen Bedürfnissen der Menschen orientieren, ihre Einschränkungen und Stärken berücksichtigen und sie bei der Erfüllung ihrer Aufgaben unterstützen.	Durchführen umfassender Bedürfnisanalysen, um die aktuellen und zukünftigen Bedürfnisse der Zielgruppen zu verstehen. Einplanen regelmässiger Dialoge und Einholen von Rückmeldungen von relevanten Stakeholdern, einschliesslich Endnutzenden, Betroffenen, Fachexpertinnen und -experten und der Zivilgesellschaft. Etablierung von Prozessen, um die Rückmeldungen adäquat in die Überarbeitung der KI-Systeme einfliessen zu lassen. Sicherstellen, dass KI-Systeme und -Anwendungen barrierefrei und benutzerfreundlich gestaltet sind, um eine breite Zugänglichkeit und Nutzung zu ermöglichen. Wo immer möglich ein alternatives, nicht-digitales Angebot bereitstellen. Durchführen von Usability-Tests mit Nutzenden aus unterschiedlichen Zielgruppen, um die Nutzungsfreundlichkeit kontinuierlich zu verbessern. Etablierung von Prozessen um die Resultate dieser Tests adäquat in die Überarbeitung der KI-Systeme einfliessen zu lassen.	Hoch

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
A6	Initialisierung, Design & Entwicklung, Betrieb & Überwachung	ZWINGEND	Die Outputs eines KI-Systems dürfen die Persönlichkeitsrechte von Einzelpersonen nicht verletzen. Ebenso können Grundrechte auch bei Entscheidungen, die nicht explizit Grundrechte betreffen, implizit betroffen sein.	Einrichten von technischen und organisatorischen Kontrollprozessen, um die Verletzung von Persönlichkeitsrechten zu verhindern. Durchführen von Folgenabschätzungen, um die grundrechtlichen Folgen auf Menschen zu bewerten. Regelmässiges Durchführen von Audits durch interne oder externe Prüfstellen.	Hoch

Tabelle 6: Anforderung KI-Systeme – Grundrechte

Ergänzende Informationen:

Die Traceability Parameter für die Anforderungen in diesem Kapitel sind wie folgt:

	Traceability		
Nr.	**Explizit**		**Implizit**
	Gesetz	***Studie/ Bericht/Int. Standard***	***Experten-Input***
A5	Bundesverfassung	Leitlinien Künstliche Intelligenz, Verhaltenscodex Datenwissenschaft Bund	-
A6	-	Leitlinien Künstliche Intelligenz, Verhaltenscodex Datenwissenschaft Bund KI-Konvention Europarat	

Tabelle 7: Anforderung KI-Systeme – Grundrechte ergänzende Informationen

7.3 Datenschutz

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
A7	Design & Entwicklung, Betrieb & Überwachung, Ausserbetriebnahme	ZWINGEND	Einhaltung der gelten Datenschutzbestimmungen (Bsp. DSG, kantonale Gesetze usw.)	1) Einhalten der Datenbearbeitungsgrundsätze, wenn Personendaten durch ein KI-System bearbeitet werden. 2) Information der betroffenen Personen über die Datenbearbeitung durch ein KI-System und über die automatisierte Einzelentscheidung.	Hoch
A8	Design & Entwicklung, Betrieb & Überwachung, Ausserbetriebnahme	ZWINGEND	Beim Einsatz von KI-Systemen besteht eine besondere Verantwortung, gegenüber den von diesen Entscheidungen betroffen Personen. Diese Verantwortung muss wahrgenommen werden. Ein hohes Risiko besteht insbesondere bei der umfangreichen Bearbeitung besonders schützenswerter Personendaten oder wenn systematisch umfangreiche öffentliche Bereiche überwacht werden.	1) Vorgängiges (vor dem Einsatz) Durchführen einer Datenschutz-Folgeabschätzung, wenn eine geplante Bearbeitung (z.B. es wird für die Bearbeitung der Personendaten ein KI-System eingesetzt) ein hohes Risiko für die Persönlichkeit oder die Grundrechte mit sich bringt. 2) Anonymisieren der Personendaten. Schulen der Mitarbeitenden, die das KI-System einsetzen. Hinweis: Eine Datenschutz-Folgenabschätzung ist insbesondere in folgenden Fällen erforderlich: ▪ systematische und umfassende Bewertung persönlicher Aspekte natürlicher Personen, einschliesslich Profiling; ▪ umfangreiche Verarbeitung sensibler Daten;	Hoch

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
				▪ systematische umfangreiche Überwachung öffentlich zugänglicher Bereiche. Diese Auflistung ist nicht abschliessend, da durch KI-Anwendungen auch ohne die Verwendung von Personendaten implizit Rückschlüsse auf Personen gemacht werden können (auch bei der Bearbeitung von “anonymisierten” Daten muss das Rückschlussrisiko berücksichtigt werden).	

Tabelle 8: Anforderung KI-Systeme – Datenschutz

Ergänzende Informationen:

Die Traceability Parameter für die Anforderungen in diesem Kapitel sind wie folgt:

	Traceability		
Nr.	**Explizit**		**Implizit**
	Gesetz	***Studie/ Bericht/Int. Standard***	***Experten-Input***
A7	Datenschutzgesetz Artikel 6 DSG Grundsätze der Datenbearbeitung, insbesondere die Prinzipien der Rechtmässigkeit, Transparenz, Verhältnismässigkeit und Zweckbindung bei der Bearbeitung von Personendaten.	-	-

	Traceability		
Nr.	**Explizit**		**Implizit**
	Gesetz	***Studie/ Bericht/Int. Standard***	***Experten-Input***
	Artikel 19 DSG Informationspflicht bei der Beschaffung von Personendaten Artikel 21 DSG Informationspflicht bei automatisierten Einzelentscheiden		
A8	Datenschutzgesetz Artikel 22 DSG Datenschutz-Folge-abschätzung	-	-

Tabelle 9: Anforderung KI-Systeme – Datenschutz ergänzende Informationen

7.4 Urheberrechte

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
A9	Initialisierung, Design & Entwicklung, Betrieb & Überwachung	ZWINGEND	KI-Systeme müssen das geistige Eigentum respektieren. Die Urheberrechte sind einzuhalten. Die unrechtmässige Verwendung geschützten Materials (z.B. Trainingsdaten, die urheberrechtlich gestützt sind) durch KI-Systeme sind zu verhindern. Die Anwendung des Outputs eines KI-Systems darf auch nicht gegen die Urheberrechte verstossen.	1) Organisatorische Massnahmen **Einhaltung Urheberrechtsgesetz** Entwickeln und Implementieren eines Prozesses, der sicherstellt, dass alle von einem KI-System bearbeiteten Inhalte auf Urheberrechtsverletzungen prüft, bevor sie genutzt oder reproduziert werden. **Audits**	Hoch

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
				Durchführen regelmässiger Audits der KI-Systeme, um die Einhaltung des Urheberrechtsgesetz bei Verwendung von Trainingsdaten und Outputs des KI-Systems, zu gewährleisten. Bei Bedarf sind korrektive Massnahmen einzuleiten. 2) Technische Massnahmen Lizenzmanagement Einrichten eines Systems zur Verwaltung und Überwachung von Lizenzen, das die Nutzung von Erzeugnissen, welche geschützt sind, steuert und dokumentiert. Content-Filtering-Systeme Einsetzen von Content-Filtering-Systemen, die urheberrechtlich geschütztes Material automatisch erkennen und blockieren. Technische Schutzmassnahmen Einbinden technischer Schutzmassnahmen, die das Kopieren und unrechtmässige Verbreiten von urheberrechtlich geschütztem Material verhindern. Hinweis: es dürfen nicht ausschliesslich technische Massnahmen eingesetzt werden.	

Tabelle 10: Anforderung KI-Systeme – Urheberrechte

Ergänzende Informationen:

Die Traceability Parameter für die Anforderungen in diesem Kapitel sind wie folgt:

	Traceability		
Nr.	**Explizit**		**Implizit**
	Gesetz	***Studie/ Bericht/Int. Standard***	***Experten-Input***
A9	Urheberrecht, Patentgesetz Artikel 2 URG Definition, was als Werk im Sinne des Urheberrechts gilt. Artikel 10 URG Verwendung des Werks. Artikel 34 PatG Lizenzerteilung	-	-

Tabelle 11: Anforderung KI-Systeme – Urheberrechte ergänzende Informationen

7.5 Autonomie

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
A10	Design & Entwicklung, Betrieb & Überwachung	ZWINGEND	KI-Systeme müssen auf die Risiken bezüglich der Autonomie der Betroffenen kontinuierlich untersucht werden.	Durchführen regelmässiger Dokumentation und Analyse der KI-Systeme sowie Prüfung weiterer Massnahmen (Bsp. Veröffentlichung der Ergebnisse). Diese Analyse sollte kontinuierlich während des Lebenszyklus des KI-Systems erstellt werden und die identifizierten Risiken berücksichtigen. In einigen Fällen kann eine Dokumentation allein ausreichend sein.	Hoch

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
A11	Design & Entwicklung, Betrieb & Überwachung	ZWINGEND	**Informationsschutz** KI-Systeme müssen Informationen, wie Geschäftsgeheimnisse, usw. vor Missbrauch schützen. **Informationssicherheit** Die Integrität, Verfügbarkeit, Vertraulichkeit und Authentizität der Daten werden durch technische und organisatorische Massnahmen gesichert.	1) Organisatorische Massnahmen: Regelmässige Audits Durchführen regelmässiger Audits der KI-Systeme, um Risiken zu identifizieren und entgegenwirken. Prüfen der Nutzungsbedingungen von KI-Systemen. ISDS-Konzept ISDS-Konzept bildet die Grundlage für die Festlegung der Massnahmen für die Informationssicherheit, gemäss HERMES. 2) Technische Massnahmen ▪ Zugangs- und Berechtigungskonzept, Authentifizierungsverfahren, Verschlüsselungsverfahren, Back-up und andere technische Massnahmen zur Informationssicherung entsprechen dem neuesten Stand der Technik	Hoch
A12	Betrieb & Überwachung	ZWINGEND	Überwachungsprozesse und -systeme müssen die Robustheit von KI-Systemen sicherstellen. .	1) Implementierung von Überwachungsprozessen: Anomalie-Erkennungssysteme Entwickeln und Einsetzen von Algorithmen zur Anomalie-Erkennung vom Output des KI-Systems. Realzeit-Monitoring	Hoch

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
				Wenn die Höhe der Risiken es erfordert, einrichten von Echtzeit-Überwachungsdiensten, die kontinuierlich Systemaktivitäten prüfen und sofort Alarm schlagen, wenn Abweichungen festgestellt werden. Berichterstattung und Protokollierung Implementieren von Protokollierungssystemen, die angemessen detaillierte Aufzeichnungen über alle Aktivitäten führen, um eine spätere Analyse zu ermöglichen und Beweise für etwaige Vorfälle zu sichern. Hinweis: nebst automatisierten Systemen sind zwingend auch organisatorische Massnahmen zu treffen, um abweichendes Verhalten manuell kennzeichnen zu können.	
A13	Design & Entwicklung, Betrieb & Überwachung	ZWINGEND	Einbindung von menschlichen Experten in die Entscheidungs- und Kontrollprozesse von KI-Systemen	Einsatz von z.B. Human-in-the-Loop Mechanismen: ▪ Überwachung und (wenn nötig) Eingriff in die Entscheidungen des KI-Systems ▪ Feedback-Schleifen, um Entscheidungen und Vorschläge des KI-Systems kontinuierlich zu verbessern ▪ KI-System eskaliert komplexe und unsichere Fälle an menschliche Experten ▪ Entscheidungsfindungsprozesse werden so gestaltet, dass Mensch	

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
				und KI-System ihre jeweiligen Stärken einbringen ▪ (Fachliche) Expertise von menschlichen Experten werden beim Training von KI-Systemen eingebunden	

Tabelle 12: Anforderung KI-Systeme – Autonomie

Ergänzende Informationen:

Die Traceability Parameter für die Anforderungen in diesem Kapitel sind wie folgt:

	Traceability		
Nr.	**Explizit**		**Implizit**
	Gesetz	***Studie/ Bericht/Int. Standard***	***Experten-Input***
A10	-	OECD-Prinzip	-
A11	-	Verhaltenscodex Datenwissenschaft Bund	-
A12	-	Verhaltenscodex Datenwissenschaft Bund	-
A13	-	-	Kernteam

Tabelle 13: Anforderung KI-Systeme – Autonomie ergänzende Informationen

7.6 Fairness

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
A14	Design & Entwicklung, Betrieb & Überwachung, Re-Evaluation	ZWINGEND	KI-Systeme dürfen keine Personen oder Personengruppen nicht auf der Grundlage bestimmter diskriminierender Merkmale benachteiligen. Gemäss art. 8 Bundesverfassung, fällt darunter die Benachteiligung wegen einer Behinderung, des Geschlechts, der Abstammung, der Sprache, der ethnischen oder sozialen Herkunft, der genetischen Merkmale, der sexuellen Ausrichtung, der Religion oder der Weltanschauung, als auch der politischen oder sonstigen Anschauung einer Person. Das KI-System muss den	1) Designentscheidungen und mögliche Kompromisse und Zielkonflikte dokumentieren. 2) Regelmässige Audits Durchführen regelmässiger Audits der KI-Systeme, um die Risiken einer Diskriminierung zu identifizieren und geeignete Massnahmen zu ergreifen, um das Risiko zu vermeiden.	Hoch
A15	Design & Entwicklung, Betrieb & Überwachung Re-Evaluation	EMPFOHLEN	Die Trainingsdaten der KI-Systeme sollten repräsentativ für alle betroffenen Personen sein und nicht nur für Teilgruppen.	1) Diversifizierte Datenerhebung Aktive Erweiterung der Datensätze, um eine Vielfalt von demografischen Gruppen und Szenarien einzuschliessen. Dies umfasst die gezielte Akquise und Einbeziehung von Daten aus unterschiedlichen Quellen, die verschiedene Geschlechter, Altersgruppen, sozioökonomische Hintergründe, geographische Regionen und weitere relevante Diversitätsdimensionen repräsentieren. 2) Bias-Analyse und -Korrektur Durchführung regelmässiger Bias-Analyse, um bestehende Verzerrungen in den Trainingsdaten zu identifizieren. Anschliessend	Mittel

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
				sollten Massnahmen zur Korrektur dieser Verzerrungen ergriffen werden, zum Beispiel durch das Anpassen von Datensätzen, Resampling-Methoden oder die Anwendung von Techniken des maschinellen Lernens, die auf Fairness und Gleichbehandlung ausgerichtet sind. Hinweis: die Auflistung der Massnahmen ist nicht abschliessend. Weitere Massnahmen können in Erwägung gezogen werden.	

Tabelle 14: Anforderung KI-Systeme – Fairness

Ergänzende Informationen:

Die Traceability Parameter für die Anforderungen in diesem Kapitel sind wie folgt:

	Traceability		
Nr.	**Explizit**		**Implizit**
	Gesetz	***Studie/ Bericht/Int. Standard***	***Experten-Input***
A14	Bundesverfassung Artikel 8	Verhaltenscodex Datenwissenschaft Bund	-
A15	-	Verhaltenscodex Datenwissenschaft Bund	Kernteam FG KI

Tabelle 15: Anforderung KI-Systeme – Fairness ergänzende Informationen

7.7 Rechenschaftspflicht

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
A16	Design & Entwicklung, Verifikation & Validation, Implementierung, Betrieb & Überwachung, Re-Evaluation, Ausserbetriebnahme	ZWINGEND	Die Verantwortlichkeiten bei KI-Systemen Eine verantwortliche Rolle (Bsp. Person, Stelle) ist für die Einhaltung der Anforderungen und Massnahmen zu definieren.	1) Erstellung eines Verantwortlichkeits-Frameworks Entwicklung eines detaillierten Frameworks, das die Verantwortlichkeiten des KI-Systems beschreibt. Dies sollte die Rollen und Zuständigkeiten für die Datenauswahl, das Modelltraining, die Validierung, die Bereitstellung und die fortlaufende Wartung abdecken. Das Framework sollte auch die Compliance mit relevanten Standards wie ISO/IEC 22989 und DIN SPEC 92001-3:2023-08 sicherstellen und Anleitungen geben, wie Verantwortlichkeiten dokumentiert und kommuniziert werden sollen. 2) Regelmässige Überprüfung und Anpassung der Verantwortlichkeiten Umsetzung von Prozessen, die eine regelmässige Überprüfung und Anpassung der Verantwortlichkeiten ermöglichen, um auf Veränderungen im KI-System oder in den relevanten Standards zu reagieren. Dies könnte über ein Compliance-Management-System geschehen, das sicherstellt, dass das KI-System und dessen Betrieb den neuesten Bestimmungen der ISO/IEC 22989	Hoch

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
				und der DIN SPEC 92001-3:2023-08 entsprechen. Dazu gehört auch die regelmässige Schulung der beteiligten Personen, um sicherzustellen, dass sie über die neuesten Anforderungen und Best Practices im Bereich KI informiert sind.	
A17	Design & Entwicklung, Verifikation & Validation, Implementierung, Betrieb & Überwachung, Re-Evaluation, Ausserbetriebnahme	ZWINGEND	Der Trainingsprozess und die Outputs von KI-Systemen müssen in allen Lebenszyklusphasen dokumentiert werden, um die Nachvollziehbarkeit sowie auch die Transparenz zu gewährleisten.	1) Versionierung und Dokumentation der Datensätze Für jede Phase der Entwicklung eines KI-Systems sollten die verwendeten Datensätze, einschliesslich ihrer Versionen und der vorgenommenen Änderungen vollständig dokumentiert werden. Dies könnte durch ein Versionskontrollsystem erfolgen, das auch die Herkunft der Daten (Data Provenance) und Änderungen an den Datensätzen festhält. Dazu gehört auch, die Methoden der Datenerhebung, -bereinigung, -anreicherung und -teilung festzuhalten. Diese Informationen sollten zwecks Transparenz möglichst öffentlich zugänglich sein. 2) Entscheidungsprotokolle und Erklärbarkeitswerkzeuge Es sollte ein System zur Protokollierung von Entscheidungen eingeführt werden, das genau festhält, wie und warum das KI-System bestimmte Entscheidungen getroffen hat. Dies kann durch den Einsatz von Erklärbar-	Hoch

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
				keitswerkzeugen (Explainable AI, XAI) ergänzt werden, die die Faktoren und Logik hinter den Vorhersagen des Modells transparent machen. Diese Werkzeuge können visuelle Darstellungen der Modellentscheidungen, Sensitivitätsanalysen oder Feature-Beitragsanalysen umfassen. Diese Informationen sollten den von Entscheidungen betroffenen auf Nachfrage in einem unkomplizierten Verfahren zur Verfügung gestellt werden. Wo immer möglich sollten diese Informationen zwecks Transparenz öffentlich zugänglich sein.	

Tabelle 16: Anforderung KI-Systeme – Rechenschaftspflicht

Ergänzende Informationen:

Die Traceability Parameter für die Anforderungen in diesem Kapitel sind wie folgt:

	Traceability		
Nr.	**Explizit**		**Implizit**
	Gesetz	***Studie/ Bericht/Int. Standard***	***Experten-Input***
A16	-	Internationaler ISO-Standard und DIN SPEC	Kernteam FG KI
A17	-	Leitlinien Künstliche Intelligenz Bund	-

Tabelle 17: Anforderung KI-Systeme – Rechenschaftspflicht ergänzende Informationen

7.8 Transparenz

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
A18	Design & Entwicklung, Verifikation & Validation, Implementierung, Betrieb & Überwachung, Re-Evaluation, Ausserbetriebnahme	ZWINGEND	Für jedes KI-System wird ein Modelltrainingsbericht vor der Implementierung erstellt. Der Bericht muss die Inhalte enthalten, die in den Massnahmen beschrieben werden. Dieser Bericht wird zwecks Transparenz öffentlich zugänglich gemacht. Der Detaillierungsgrad des Berichts kann je nach Kontext des KI-Systems sich variieren.	1) Inhalte des Modelltrainingsberichts **Einführung und Zweck** Eine kurze Beschreibung des zu lösenden Problems und warum ein KI-Modell zur Lösung in Betracht gezogen wurde und als die beste Lösung identifiziert wurde. **Datenbeschreibung** ▪ Herkunft der Daten ▪ Datenstruktur und -format ▪ Statistische Zusammenfassungen (z. B. Mittelwert, Median, Varianz) ▪ Anzahl der Datenpunkte, fehlende Werte, etc. ▪ **Vorverarbeitung der Daten** ▪ Methoden zur Datenbereinigung ▪ Feature-Engineering und -Auswahl ▪ Normalisierung und Skalierung ▪ Aufteilung in Trainings-, Validierungs- und Testdaten **Modellauswahl** ▪ Beschreibung der verschiedenen in Betracht gezogenen Modelle (z. B.	Hoch

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
				neuronale Netze, Entscheidungs-bäume, SVMs) ▪ Begründung für die Auswahl eines bestimmten Modells **Trainingsprozess** ▪ Trainingsstrategie und -parameter (z. B. Lernrate, Batch-Grösse) ▪ Verwendete Optimierungsalgorith-men ▪ Regularisierungsmethoden, falls verwendet **Ergebnisse und Performance-Metriken** ▪ Trainings- und Validierungsverlust über die Zeit ▪ Weitere Metriken je nach Anwen-dungsfall (z. B. Genauigkeit, F1-Score, AUC) ▪ Eventuelle Überanpassung (Overfit-ting) und Massnahmen dagegen **Visualisierungen** ▪ Lernkurven ▪ Eventuell Gewichts- und Aktivie-rungsdistributionen ▪ Für Klassifikationsaufgaben: Konfu-sionsmatrizen, ROC-Kurven usw.	

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
				Diskussion und Interpretation ▪ Analyse der Ergebnisse und des Modellverhaltens ▪ Potenzielle Schwächen des Modells und Vorschläge zur Verbesserung **Abschluss und Empfehlungen** ▪ Zusammenfassung der wichtigsten Erkenntnisse ▪ Empfehlungen für die nächste Iteration oder Implementierung **Anhänge** Code-Schnipsel oder Verweis auf den vollständigen Code.	

Tabelle 18: Anforderung KI-Systeme – Transparenz

Ergänzende Informationen:

Die Traceability Parameter für die Anforderungen in diesem Kapitel sind wie folgt:

	Traceability		
Nr.	**Explizit**		**Implizit**
	Gesetz	***Studie/ Bericht/Int. Standard***	***Experten-Input***
A18	-	Leitlinien Künstliche Intelligenz Bund (Transparenz)	Kernteam FG KI

Tabelle 19: Anforderung KI-Systeme – Transparenz ergänzende Informationen

7.9 Erklärbarkeit

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
A19	Design & Entwicklung	ZWINGEND	Wenn ein KI-System Risiken für die Persönlichkeit und die Grundrechte für die Endnutzer bzw. Betroffene birgt, dann sollten die KI-Systeme über eingebaute Erklärbarkeitsfunktionen verfügen, die Logik und Grundlagen einer Entscheidung verständlich und klar darlegen können, sowohl für Experten als auch für Laien, unterstützt durch entsprechende Trainings. Dazu gehören auch geeignete Weiterbildungsangebote für Stakeholder, abgestimmt auf ihre jeweilige Rolle.	1) Implementierung von Explainable AI (XAI) Methoden: Integration von Erklärungsmethoden und -Werkzeuge in die Algorithmen, die die Entscheidungsfindung in einfacher Sprache zusammenfassen können. Diese Werkzeuge könnten beispielsweise die wichtigsten Merkmale hervorheben, die zu einer Entscheidung geführt haben, oder eine Schritt-für-Schritt-Nachverfolgung der Entscheidungswege des Algorithmus anbieten. 2) Benutzerfreundliche Visualisierung von Daten und Entscheidungen Entwicklung von interaktiven Visualisierungen, die es Benutzern ermöglichen, die Daten und die darauf basierenden Entscheidungen des Algorithmus nachzuvollziehen. Zum Beispiel könnten Diagramme, Flussdiagramme oder andere grafische Elemente eingesetzt werden, um die Logik hinter den Entscheidungen zu veranschaulichen und diese auch für Benutzer ohne technischen Hintergrund greifbar zu machen. 1) Weiterbildungsangebote und Trainings zu KI-Systemen	Hoch

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
				Auf die Stakeholder und ihre Rollen abgestimmte Weiterbildungsangebote zu KI und KI-Systemen. Hinweis: Die Auflistung der oben genannten Massnahmen ist nicht abschliessend und je nach Kontext des KI-System kann geprüft werden, ob und welche diese Massnahmen erforderlich sind.	
A20	Design & Entwicklung	EMPFHOLEN	Die Funktionsweise von KI-Systemen sollte möglichst transparent aufgezeigt und erklärt werden.	1) Bereitstellung von Tools oder Schnittstellen, die Einblicke in die Funktionsweise des KI-Systems bieten. 2) System für Nutzeranfragen Einsatz von Mechanismen oder Systemen um Nutzeranfragen zu Entscheidungsprozessen von KI-Systemen effektiv und nutzerorientiert zu adressieren. Nutzern können z.B. statistische Daten, Verarbeitungsabläufe und Entscheidungspfade von KI-Systemen angezeigt werden. 3) Dokumentation und Hilfsmittel für die Entscheidungsfindung Erstellung einer Dokumentation und Hilfsmittel, die erklären, wie das KI-System zu seinen Ergebnissen kommt. Dazu können Whitepapers, FAQs oder Tutorial-Videos gehören, die die Technologie hinter dem System in leicht verständlicher Form erklären.	Mittel

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
A21	Design & Entwicklung Verifikation & Validation	OPTIONAL	Um die Erklärbarkeit der KI-Systeme aufzuweisen, wird zunächst beurteilt, ob es sich bei einem KI-System um ein Black-Box oder White-Box handelt. Hierfür sollen erprobte Ansätze eingesetzt werden.	Als Massnahme wird z.B. der Ansatz von Arrieta et al. 2019 empfohlen. Dieser Ansatz beinhaltet Kriterien, die für die Erklärbarkeit von KI-Systemen gewährleisten und helfen dabei, das Vertrauen in solche Systeme zu erhöhen. Bei den Kriterien aus dem oben genannten Ansatz handelt es sich um solche wie die Simulierbarkeit, die Unterteilbarkeit und die algorithmische Transparenz, welche wie folgt beschrieben werden: **Simulierbarkeit** meint, dass alle Rechenschritte eines KI-Modells in angemessener Zeit auswertbar sein müssen. Das gilt eben beispielsweise nicht mehr für Neuronale Netze, weil sie Millionen an Gewichten, sprich: Rechenoperationen, umfassen. **Unterteilbarkeit** bedeutet, dass alle Bestandteile eines KI-Modells wie Daten, Parameter und Berechnungen intuitiv sein müssen. Auch dies gilt nur für wenige KI-Modelle wie beispielsweise Entscheidungsbäume, die anhand klarer Kriterien Entscheidungen treffen. **«Algorithmische Transparenz»** schliesslich sollte der Lernalgorithmus selbst auch nachvollziehbar sein. Bei einem linearen Modell ist dies möglich, weil es eine nach-	Tief

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
				vollziehbare Entscheidungsgrenze gibt (etwas gehört zur Menge a oder zur Menge b). Im Vergleich dazu ist die Entscheidungsgrenze bei Neuronalen Netzen stark nichtlinear und nicht mehr explizit gegeben. Die Beurteilung für den Nachweis, ob es sich bei einem KI-System um ein Black-Box oder White-Box handelt, kann mit der Zuordnung erfolgen, wie es in der Abbildung 4 im Anhang beschrieben ist.	
A22	Design & Entwicklung Verifikation & Validation	OPTIONAL	Für die Erklärbarkeit von einem KI-System ist ein oder mehrere Frameworks einzusetzen.	1) Einsatz von Frameworks für die Erklärbarkeit Bei der Auswahl eines Frameworks zur Erklärbarkeit von KI und ist nicht nur die Technik, sondern auch die jeweiligen Anforderungen und das Publikum zu berücksichtigen. Es ist wichtig, das richtige Werkzeug für den spezifischen Kontext und die spezifischen Bedürfnisse zu wählen, siehe Tabelle im Anhang. Beispiele für Frameworks ▪ SHAP (SHapley Additive exPlanations) ▪ LIME (Local Interpretable Model-Agnostic Explanations) ▪ What-if Tool ▪ AIX360 (AI Explainability 360) ▪ Skater	Tief

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
A23	Design & Entwicklung Verifikation & Validation	OPTIONAL	Das erklärbare KI-Framework sollte die Möglichkeit zur Visualisierung von Ergebnissen anbieten.	1) Entwicklung eines integrierten Visualisierungstools Implementierung von Tools direkt in das Framework, die automatisch Visualisierungen für die Ergebnisse der KI generieren. Dies könnte interaktive Graphen, Heatmaps oder Baumstrukturen umfassen, die zeigen, wie Eingaben zu Entscheidungen geführt haben. 2) Schnittstellen zu bestehenden Visualisierungsbibliotheken Schaffung von Schnittstellen zu etablierten Datenvisualisierungsbibliotheken wie Matplotlib, Seaborn oder D3.js. Dadurch können Nutzer des Frameworks die leistungsstarken Visualisierungswerkzeuge nutzen, um die Ergebnisse und Entscheidungen der KI-Modelle darzustellen und zu interpretieren.	Tief
A24	Design & Entwicklung Verifikation & Validation	OPTIONAL	Das erklärbare KI-Framework sollte verschiedener Datentypen (z. B. Tabellendaten, Text, Bilder) unterstützen.	1) Modularisierung des Frameworks Entwicklung von einem Framework, dass es modulare Komponenten enthält, die jeweils auf die Verarbeitung und Analyse eines bestimmten Datentyps spezialisiert sind. Jedes Modul kann dann spezifische Algorithmen und Verarbeitungsmethoden für den jeweiligen Datentyp implementieren, um dessen Besonderheiten und Anforderungen gerecht zu werden.	Tief

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
				2) Erweiterbare Schnittstellen für Dateneingabe Sicherstellung, dass das Framework erweiterbare Schnittstellen für die Eingabe von Daten hat. Diese Schnittstellen sollten so gestaltet sein, dass sie leicht um neue Datentypen erweitert werden können, sei es durch das Hinzufügen von Code für neue Formate oder durch die Integration von Plugins oder Bibliotheken, die speziell für diese Daten entwickelt wurden.	
A25	Design & Entwicklung Verifikation & Validation	OPTIONAL	Das erklärbare KI-Framework sollte die Option zur Interaktion mit dem Modell zur weiteren Untersuchung anbieten.	1) Interaktive Query-Schnittstelle Implementierung von einer benutzerfreundlichen Query-Schnittstelle, die es Benutzern ermöglicht, Anfragen an das Modell zu stellen und verschiedene Szenarien zu testen. Diese Schnittstelle könnte beispielsweise Benutzerinputs akzeptieren, um Vorhersagen zu generieren, und gleichzeitig erklären, wie verschiedene Eingabewerte die Ergebnisse beeinflussen. 2) What-if-Analyse-Tools Tools zur Verfügung stellen, die What-if-Analysen unterstützen. Diese erlauben es Benutzern, die Auswirkungen von Veränderungen in den Eingabedaten auf die Modellausgaben zu sehen. Benutzer könnten mit den Eingabewerten experimentieren, um zu	Tief

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
				verstehen, wie robust das Modell auf Änderungen reagiert und welche Faktoren die stärkste Auswirkung auf die Vorhersagen haben.	
A26	Design & Entwicklung Verifikation & Validation	OPTIONAL	Das erklärbare KI-Framework sollte Datenschutz und Anonymität bei der Datenanalyse berücksichtigen.	1) Implementierung von Datenanonymisierungstechniken und privacy enhancing technologies bei Bedarf Anonymisierung von Personendaten vor der Analyse; Dies kann durch Techniken wie das Maskieren, Hashing oder das K-Durchschnitts-Verfahren geschehen, um sicherzustellen, dass die betroffene Person nicht identifiziert werden kann. 2) Zugriffssteuerung und Audit-Trails Festlegung von Zugriffskontrollen, um sicherzustellen, dass nur autorisierte Nutzer Zugang zu sensiblen Daten haben. Führen Sie Audit-Trails, die alle Zugriffe und Aktionen mit den Daten aufzeichnen, um eine nachvollziehbare Historie von Datenzugriffen und -veränderungen zu gewährleisten.	Tief
A27	Design & Entwicklung Verifikation & Validation	OPTIONAL	Die Erklärungsmethoden für KI-Systeme sollten validiert und in der Literatur gut dokumentiert sein.	1) Schaffung von Best-Practice-Richtlinien Erarbeitung von Best-Practice-Richtlinien für die Anwendung und Dokumentation von Erklärungsmethoden. Diese sollten auf aktueller Forschung basieren und praktische An-	Tief

Nr.	KI-Lebenszyklus	Notation	Anforderung	Massnahme/n	Klassifikation
				leitungen zur Implementierung und Überprüfung von Erklärungsmethoden enthalten. Diese Richtlinien könnten in Form von technischen Berichten, Online-Ressourcen oder in Workshops und Schulungen verbreitet werden. 2) Peer-Review und Publikation Durchführung eines Peer-Review-Verfahren, um die Erklärungsmethoden von unabhängigen Experten begutachten zu lassen. Anschliessend sollten die Methoden zusammen mit Validierungsergebnissen in wissenschaftlichen Fachzeitschriften oder auf Konferenzen veröffentlicht werden, um die Methoden einer breiten wissenschaftlichen Gemeinschaft zugänglich zu machen.	

Tabelle 20: Anforderung KI-Systeme – Erklärbarkeit

Ergänzende Informationen:

Die Traceability Parameter für die Anforderungen in diesem Kapitel sind wie folgt:

	Traceability		
Nr.	**Explizit**		**Implizit**
	Gesetz	***Studie/ Bericht/Int. Standard***	***Experten-Input***
A19	-	-	Kernteam FG KI
A20	-	-	Kernteam FG KI

	Traceability		
Nr.	**Explizit**		**Implizit**
	Gesetz	***Studie/ Bericht/Int. Standard***	***Experten-Input***
A21	-	Publikation: Ansatz von Arrieta et al. 2019	-
A22	-	Publikation: Ansatz von Arrieta et al. 2019	-
A23	-	-	Kernteam FG KI
A24	-	-	Kernteam FG KI
A25	-	-	Kernteam FG KI
A26	-	-	Kernteam FG KI
A27	-	-	Kernteam FG KI

Tabelle 21: Anforderung KI-Systeme – Erklärbarkeit ergänzende Informationen

8 Praxistauglichkeit KI-Standards

Beim vorliegenden Standard handelt es sich um einen neuen Standard, der noch nicht im Einsatz ist. Deshalb werden die Anforderungen, die in diesem Standard definiert werden, anhand von Prototypen bzw. von Tools geprüft bzw. getestet, um die Praxistauglichkeit zu prüfen. Diese Prüfung ist nicht Bestand dieses Standards, weshalb diese Bestimmung hier nur als Hinweis gilt.

9 Sicherheitsüberlegungen

Es sind keine besonderen Sicherheitsüberlegungen vorhanden.

10 Haftungsausschluss/Hinweise auf Rechte Dritter

eCH-Standards, welche der Verein **eCH** den Benutzenden zur unentgeltlichen Nutzung zur Verfügung stellen oder welche **eCH** referenzieren, haben nur den Status von Empfehlungen. Der Verein **eCH** haftet in keinem Fall für Entscheidungen oder Massnahmen, welche den Benutzenden auf Grund dieser Dokumente trifft und / oder ergreift. Die Benutzenden sind verpflichtet, die Dokumente vor deren Nutzung selbst zu überprüfen und sich gegebenenfalls beraten zu lassen. **eCH**-Standards können und sollen die technische, organisatorische oder juristische Beratung im konkreten Einzelfall nicht ersetzen.

In **eCH**-Standards referenzierte Dokumente, Verfahren, Methoden, Produkte und Standards sind unter Umständen markenrechtlich, urheberrechtlich oder patentrechtlich geschützt. Es liegt in der ausschliesslichen Verantwortlichkeit der Benutzenden, sich die allenfalls erforderlichen Rechte bei den jeweils berechtigten Personen und/oder Organisationen zu beschaffen.

Obwohl der Verein **eCH** all seine Sorgfalt darauf verwendet, die **eCH**-Standards sorgfältig auszuarbeiten, kann keine Zusicherung oder Garantie auf Aktualität, Vollständigkeit, Richtigkeit bzw. Fehlerfreiheit der zur Verfügung gestellten Informationen und Dokumente gegeben werden. Der Inhalt von **eCH**-Standards kann jederzeit und ohne Ankündigung geändert werden.

Jede Haftung für Schäden, welche den Benutzenden aus dem Gebrauch der **eCH**-Standards entstehen ist, soweit gesetzlich zulässig, wegbedungen.

11 Urheberrechte

Wer **eCH**-Standards erarbeitet, behält das geistige Eigentum an diesen. Allerdings verpflichten sich die Erarbeitenden, ihr betreffendes geistiges Eigentum oder ihre Rechte an geistigem Eigentum anderer, sofern möglich, den jeweiligen Fachgruppen und dem Verein **eCH** kostenlos zur uneingeschränkten Nutzung und Weiterentwicklung im Rahmen des Vereinszweckes zur Verfügung zu stellen.

Die von den Fachgruppen erarbeiteten Standards können unter Nennung der jeweiligen urhebenden Person von **eCH** unentgeltlich und uneingeschränkt genutzt, weiterverbreitet und weiterentwickelt werden.

eCH-Standards sind vollständig dokumentiert und frei von lizenz- und/oder patentrechtlichen Einschränkungen. Die dazugehörige Dokumentation kann unentgeltlich bezogen werden.

Diese Bestimmungen gelten ausschliesslich für die von **eCH** erarbeiteten Standards, nicht jedoch für Standards oder Produkte Dritter, auf welche in den **eCH**-Standards Bezug genommen wird. Die Standards enthalten die entsprechenden Hinweise auf die Rechte Dritter.

Anhang A – Referenzen & Bibliographie

Übersicht White-Box/Blackbox-Charakter von Modellen

Übersicht über White-Box-/Black-Box-Charakter von Modellen, die für das maschinelle Lernen eingesetzt werden

KI-Modell	Transparenz			White-Box/ Black-Box	Post-hoc-Analyse notwendig?
	Simulierbarkeit	Unterteilbarkeit	Algorith. Transparenz		
Neuronale Netze	X	X	X	Black-Box	Notwendig: Werkzeuge in Kap. 3
Ensemble-Modelle (z. B. Tree Ensembles)	X	X	X	Black-Box	Notwendig: Werkzeuge in Kap. 3
Support Vector Machines	X	X	X	Black-Box	Notwendig: Werkzeuge in Kap. 3
Bayes-Netze	(√)	(√)	√	White-Box*	Nicht notwendig
Lineare/logistische Regressionsmodelle	(√)	(√)	√	White-Box*	Nicht notwendig
Entscheidungsbäume (Decision Trees)	(√)	(√)	√	White-Box*	Nicht notwendig

Abbildung 3: Übersicht über White-Box-/Black-Box-Charakter von Modellen, die für das maschinelle Lernen eingesetzt werden in Anlehnung Arrieta, et., 2019)

Framework-Übersicht

Framework	Beschreibung	Nutzen
SHAP (SHapley Additive exPlanations)	Nutzt Spieltheorie, um den Einfluss jeder Feature auf die Vorhersage in einem kohärenten und theoretisch fundierten Rahmen zu erklären.	Dieses Framework ermöglicht detaillierte Einblicke in die Beiträge einzelner Features und bietet konsistente Erklärungen.
LIME (Local Interpretable Model-Agnostic Explanations)	Erstellt lokale approximative Modelle, um Erklärungen für individuelle Vorhersagen zu liefern.	Dieses Framework kann mit jedem Modell verwendet werden und stellt visuelle Erklärungen bereit, um herauszufinden, warum ein bestimmtes Modell zu einer bestimmten Entscheidung gekommen ist.
What-if Tool	Ein visuelles Interface für ML-Modelle, das es Benutzern ermöglicht, die Auswirkungen von Änderungen an den Eingabedaten oder dem Modell selbst zu sehen.	Dieses Framework fördert das Verständnis und die Intuition über Modelle und ihre Daten.

AIX360 (AI Explainability 360):	Eine umfangreiche Suite von Algorithmen zur Erklärbarkeit und Interpretierbarkeit von IBM Research.	bietet verschiedene Techniken für unterschiedliche Anwendungsfälle und ermöglicht so eine breite Abdeckung von Erklärbarkeitsbedürfnissen.
Skater:	Ein einheitliches Framework zur Interpretation von Modellvorhersagen und zur Erklärung der Funktionsweise von komplexen Modellen.	vereinfacht den Prozess der Modellinterpretation und ermöglicht die Visualisierung von Feature-Einflüssen und -Wichtigkeiten.

Abbildung 4: Framework-Übersicht, eigene Darstellung in Anlehnung Arrieta, et., 2019)

Anhang B – Mitarbeit & Überprüfung

Name	**Organisation**
Anna Mätzener	Kanton Zürich
Heidi Ates	-
Mark Strauch	Stadt Zürich
Mevlüt Polat	ETH
Robin Pekerman	Kanton Zürich
Ursulina Kölbener	Kanton Appenzell I.Rh.I

Das BAKOM und die BFS wurden konsultiert.

Anhang C – Abkürzungen und Glossar

Das Glossar kommt in diesem Dokument im Kapitel 3 unter Terminologie vor.

Anhang D – Änderungen gegenüber Vorversion

Dies ist die erste Version.

Anhang E – Abbildungsverzeichnis

Anhang F – Tabellenverzeichnis