Übersicht: Besonderheiten der Risikobewertung bei KI-Systemen

Risikobereich	Beschreibung/typische Ausprägung	Mögliche technische und organisatorische Massnahmen (TOM)
Intransparente Entscheidungslogik („Blackbox“)	Die Funktionsweise des KI-Systems ist für Verantwortliche und betroffene Personen nur eingeschränkt nachvollziehbar. Es bleibt unklar, welche Faktoren konkret zu einem bestimmten Ergebnis geführt haben, wie einzelne Eingaben gewichtet wurden und welche Entscheidungslogik oder Modellarchitektur zugrunde liegt. Dies erschwert sowohl die interne Kontrolle als auch die rechtliche Bewertung der Bearbeitung.	Einsatz erklärbarer Modelle (Explainable AI), detaillierte Dokumentation von Modellen und Entscheidungslogik, interne Prüf- und Freigabeprozesse sowie regelmässige Audits
Fehlende Erklärbarkeit gegenüber betroffenen Personen	Entscheidungen oder Ergebnisse können gegenüber betroffenen Personen nicht verständlich oder nachvollziehbar begründet werden. Dies ist insbesondere bei automatisierten Entscheidungen problematisch, da betroffenen Personen ihre Rechte (z. B. Widerspruch oder Überprüfung) nicht effektiv wahrnehmen können.	Entwicklung von Transparenzkonzepten, verständliche und adressatengerechte Erläuterungen, Schulung von Mitarbeitenden im Umgang mit erklärungsbedürftigen KI-Ergebnissen
Verzerrte Trainingsdaten	Trainingsdaten enthalten bestehende Ungleichheiten, Vorurteile oder strukturelle Verzerrungen, die vom Modell übernommen und teilweise verstärkt werden. Diese Verzerrungen sind häufig nicht offensichtlich und können sich erst im Betrieb zeigen.	Systematische Prüfung und Bereinigung von Trainingsdaten, Diversitäts- und Qualitätschecks, regelmässige Modellvalidierung sowie Dokumentation der Datengrundlage
Diskriminierungsrisiken (Bias)	Das System führt zu systematischen Benachteiligungen oder Bevorzugungen bestimmter Personengruppen, etwa aufgrund von Geschlecht, Alter, Herkunft oder sozioökonomischen Faktoren, auch ohne explizite Programmierung solcher Effekte.	Durchführung strukturierter Bias-Tests, kontinuierliches Monitoring auf Diskriminierung, Anpassung von Modellen, Schwellenwerten und Entscheidungsregeln
Unzutreffende oder fehlerhafte Ergebnisse („Halluzinationen“)	KI-Systeme erzeugen plausibel wirkende, aber sachlich falsche oder irreführende Inhalte und Bewertungen, die ohne weitere Prüfung übernommen werden können und zu erheblichen Fehlentscheidungen führen.	Implementierung von Plausibilitätsprüfungen, Vier-Augen-Prinzip, verpflichtende menschliche Endkontrolle sowie klare Freigabeprozesse
Dynamische Systemveränderungen („Model Drift“)	Das Verhalten des Systems verändert sich im Zeitverlauf durch neue Daten, veränderte Nutzung oder Nachtrainingsprozesse, sodass Ergebnisse nicht mehr stabil, konsistent oder vorhersehbar sind. Risiken entstehen häufig schleichend.	Regelmässige Re-Validierung und Re-Training unter kontrollierten Bedingungen, kontinuierliches Monitoring, Versionierung und Dokumentation von Modellständen
Unklare Datenflüsse	Es ist nicht vollständig transparent, welche Daten wann, wie und durch wen bearbeitet, gespeichert oder weitergegeben werden. Besonders bei externen KI-Diensten oder komplexen Systemlandschaften besteht ein erhöhtes Risiko intransparenter Datenbearbeitung. Häufig fehlen zudem klare Informationen zu Datenflüssen, Speicherorten und beteiligten Dritten, sodass eine vollständige Nachvollziehbarkeit nur eingeschränkt möglich ist.	Durchführung detaillierter Datenflussanalysen, Abschluss und Kontrolle von AV-Verträgen, klare Beschränkung und Dokumentation von Datenübermittlungen
Unklare Herkunft der Trainingsdaten	Die Herkunft, Qualität und rechtliche Zulässigkeit der Trainingsdaten ist nicht eindeutig nachvollziehbar oder dokumentiert. Dies kann zu Verstössen gegen Datenschutz- oder Urheberrecht führen.	Sorgfältige Anbieterprüfung (Due Diligence), vertragliche Zusicherungen zur Datenherkunft, umfassende Dokumentation der verwendeten Datensätze
Umfangreiche Datenbearbeitung	Es werden grosse Mengen Personendaten bearbeitet oder analysiert, häufig auch in aggregierter Form, wodurch das Risiko für die Rechte und Freiheiten der betroffenen Personen deutlich erhöht wird.	Umsetzung von Datenminimierung, Pseudonymisierung und Zweckbindung, Beschränkung auf erforderliche Daten sowie Zugriffsbeschränkungen
Zusammenführung mehrerer Datenquellen	Daten aus unterschiedlichen Quellen werden kombiniert, wodurch umfassende Profile entstehen und neue, teils sensible Rückschlüsse möglich werden, die ursprünglich nicht beabsichtigt waren.	Sicherstellung der Zweckbindung, Begrenzung der Datenzusammenführung, differenzierte Zugriffskonzepte und regelmässige Überprüfung der Datenintegration

Automatisierte Entscheidungsfindung	Entscheidungen werden ganz oder teilweise automatisiert auf Basis von KI-Ergebnissen getroffen, ohne dass eine ausreichende menschliche Kontrolle oder Plausibilisierung erfolgt.	Implementierung von „Human in the Loop"-Ansätzen, klare Entscheidungsprozesse, definierte Eingriffs- und Übersteuerungsmöglichkeiten
Nur formale menschliche Kontrolle	Eine vorgesehene menschliche Kontrolle findet faktisch nicht statt, da Ergebnisse aus Effizienzgründen routinemässig ungeprüft übernommen werden („Automation Bias").	Schulung der Mitarbeitenden, verbindliche Prüfprozesse, klare Verantwortlichkeiten und Stichprobenkontrollen
Eingeschränkte Betroffenenrechte	Auskunfts-, Lösch- oder Widerspruchsrechte können aufgrund fehlender Transparenz, komplexer Systeme oder technischer Einschränkungen nicht effektiv umgesetzt werden.	Etablierung klarer Prozesse zur Wahrung von Betroffenenrechten, Verbesserung der Systemtransparenz und Dokumentation
Zweckänderung der Datenbearbeitung	Daten werden für andere Zwecke verwendet als ursprünglich vorgesehen, etwa zur weiter Bearbeitung als Trainingsdaten oder für neue Analysen, ohne ausreichende Rechtsgrundlage.	Klare Zweckdefinition, getrennte Bearbeitung, regelmässige Prüfung und Dokumentation der Rechtsgrundlagen
Sicherheitsrisiken (z. B. Data Poisoning)	Systeme können durch gezielte Manipulation von Trainingsdaten oder Modellen beeinflusst werden, was zu fehlerhaften oder verzerrten Ergebnissen führt.	Umsetzung technischer Sicherheitsmassnahmen, Zugriffskontrollen, Monitoring und Schutzmechanismen gegen Manipulation
Abhängigkeit von externen Anbietern	KI-Systeme werden über externe Dienstleister betrieben, wodurch die Kontrolle über Datenbearbeitung, Sicherheitsstandards und Modellverhalten eingeschränkt ist.	Abschluss und Kontrolle von AV-Verträgen, Durchführung von Anbieter-Audits, Auswahl vertrauenswürdiger Anbieter
Drittlandübermittlungen	Daten werden bei Nutzung von KI-Diensten in Drittländer übertragen, ggf. ohne angemessenes Datenschutzniveau oder ausreichende Garantien.	Einsatz von Standardvertragsklauseln, Durchführung von Transfer Impact Assessments, bevorzugt Nutzung von EU-basiertem Hosting
Fehlende Dokumentation	Bearbeitung, Entscheidungslogik und Datenflüsse sind nicht ausreichend dokumentiert, was die Nachvollziehbarkeit und Rechenschaftspflicht erheblich erschwert.	Durchführung und Pflege einer DSFA, Verzeichnis von Bearbeitungstätigkeiten, interne Richtlinien und Dokumentationspflichten
Risiken im laufenden Betrieb	Risiken entstehen oder verändern sich erst im praktischen Einsatz, etwa durch neue Daten, veränderte Nutzung oder externe Einflüsse.	Laufendes Monitoring, regelmässige Überprüfung und Aktualisierung der DSFA sowie kontinuierliche Risikoanalyse
Memorization (Wiedergabe von Trainingsdaten)	Das System kann Inhalte aus Trainingsdaten reproduzieren, einschliesslich Personendaten oder sensibler Informationen, ohne dass dies beabsichtigt ist.	Einsatz von Filtermechanismen, Auswahl geprüfter Modelle, konsequente Datenminimierung und Zugriffsbeschränkungen
Overreliance / Automatisierungsbias	Nutzer verlassen sich zu stark auf KI-Ergebnisse und hinterfragen diese nicht mehr kritisch, was zu Fehlentscheidungen und Kontrollverlust führen kann.	Schulungen, klare Nutzungsvorgaben, verpflichtende kritische Prüfung und Dokumentation von Entscheidungen
Fehlende Zweckkontrolle bei generativer KI	Eingaben können ungewollt für Trainings- oder andere Zwecke verwendet werden, insbesondere bei offenen oder cloudbasierten Systemen.	Einschränkung von Eingaben, Nutzungsdaten schutzkonformer Systeme, klare Richtlinien zur Verwendung
Unkontrollierte Nutzung („Shadow AI")	Mitarbeitende nutzen KI-Tools ohne Freigabe oder klare Vorgaben, wodurch Kontroll-, Sicherheits- und Compliance-Risiken entstehen.	Einführung verbindlicher KI-Richtlinien, Freigabeprozesse, Schulungen und Sensibilisierung
Prompt-Leakage / Datenabfluss	Sensible Informationen werden über Eingaben (Prompts) in externe Systeme übertragen und dort gespeichert oder weiterbearbeitet.	Richtlinien zur Dateneingabe, technische Sperren, Sensibilisierung der Mitarbeitenden
Fehlende Lösch- und Kontrollmöglichkeiten	Einmal bearbeitete Daten lassen sich nicht oder nur eingeschränkt löschen oder kontrollieren, insbesondere bei Modelltraining oder externen Systemen.	Anbieterprüfung, vertragliche Regelungen, konsequente Datensparsamkeit und Datenkontrolle
Geringe Robustheit	Systeme reagieren empfindlich auf ungewöhnliche, fehlerhafte oder manipulierte Eingaben und liefern dadurch unzuverlässige oder inkonsistente Ergebnisse.	Durchführung von Robustheitstests, Monitoring, Absicherung gegen Missbrauch und Fehlverhalten
Fehlende Governance Strukturen	Es bestehen keine klaren Zuständigkeiten, Prozesse oder Verantwortlichkeiten für den Einsatz, die Überwachung und die Weiterentwicklung von KI-Systemen im Unternehmen.	Einführung von KI-Governance-Strukturen, klare Verantwortlichkeiten, etablierte Kontroll- und Freigabemechanismen