WHITEPAPER

KI-Governance

Warum Schatten-KI ein Steuerungsbefund ist und Governance den kontrollierten Einsatz erst ermöglicht

Carsten Beck

Cybersecurity Auditor | ISO/IEC 27001 & ISO 22301 Lead Auditor

MAS Information Systems Management

Erstveröffentlichung September 2026 · Version 1.1 (Oktober 2026)

Executive Briefing

Der Einsatz von künstlicher Intelligenz hat sich in vielen Organisationen schneller verbreitet, als die dazugehörige Steuerung aufgebaut wurde. Das Ergebnis ist Schatten-KI: Nutzung, die stattfindet, ohne bewusst entschieden, dokumentiert oder verantwortet zu sein.

Schatten-KI ist dabei selten primär ein Technologie- oder Disziplinproblem der Mitarbeitenden. Sie ist vor allem ein Steuerungsbefund. Sie zeigt, wo die Organisation nicht weiss, welche Informationen in welche KI-Bearbeitung gelangen, und wo der freigegebene Weg den legitimen Bedarf nicht deckt. Ein Messwert ist Schatten-KI nicht. Wer sie antrifft, erfährt aber weniger über die KI-Kompetenz der Belegschaft als über Lücken in der eigenen Steuerung.

Pauschale Verbote und die Überwachung von Mitarbeitenden helfen dabei wenig. Ein Verbot ohne praktikable Alternative kann die Nutzung auf private Geräte oder andere nicht kontrollierte Kanäle verlagern. Aus einem steuerbaren Risiko wird so ein weitgehend unsichtbares. Wirksam sind Transparenz über die tatsächliche Nutzung, klare Regeln, eine benannte Verantwortung, ein freigegebener Weg, der den legitimen Bedarf abdeckt, und für Funktionen, die Anbieter selbst aktivieren, eine vorbereitete Lieferantensteuerung. Governance ist dabei kein Selbstzweck, sondern Voraussetzung für kontrollierte Entwicklungsfähigkeit: Der Weg von individueller assistiver Nutzung bis zu agentischen Systemen[1] verändert die Art des Risikos, und mit ihm die Anforderungen an die Steuerung.

Gegen den Aufbau einer Steuerung wird häufig eingewendet, die rechtliche Lage sei noch offen. Das trifft so nicht zu. Auch ohne übergreifendes KI-Gesetz wird der Einsatz von KI in der Schweiz bereits heute durch bestehendes Datenschutz- und Arbeitsrecht, durch Geheimhaltungspflichten und je nach Branche durch Aufsichtsrecht erfasst.

Dieses Whitepaper beschreibt KI-Governance als Steuerungsaufgabe: wie Schatten-KI entsteht, wozu Governance dient, welche Stufen der KI-Nutzung zu unterscheiden sind, welche Elemente eine KI-Governance umfasst, in welcher Reihenfolge sie aufgebaut werden und woran sich ihre Wirksamkeit zeigt. Angesprochen sind Leitungs- und Aufsichtsgremien, fachliche Leistungsverantwortliche sowie Verantwortliche

[1] Agentische Systeme: KI-Systeme, die nicht nur Inhalte erzeugen, sondern auf Basis eines Ziels eigenständig Handlungen in angebundenen Systemen ausführen, etwa Daten ändern, Nachrichten versenden oder Prozesse auslösen.

in IT, Informationssicherheit, Datenschutz, Risiko, Beschaffung, Compliance und Audit. Kapitel 7 beschreibt die Prüfnachweise, Kapitel 8 enthält Leitfragen für Leitungs- und Aufsichtsgremien.

1 Ausgangslage: KI als unsichtbares Steuerungsobjekt

Schatten-IT und frei verfügbare Cloud-Dienste kennen Organisationen seit Langem. KI verschärft diese Dynamik in einem Punkt, der für die Steuerung entscheidend ist. Ihre Einführung erfordert weder ein Projekt noch eine Beschaffung noch eine Installation. Eine Anmeldung im Browser genügt. Oder es genügt, dass ein bestehender Anbieter eine KI-Funktion aktiviert.

Hinzu kommt die Geschwindigkeit. Frühere Technologiewellen benötigten Jahre, bis Anwendungen ausgereift, Preise tragbar und Mitarbeitende geschult waren. KI-Dienste laufen auf bereits vorhandenen Geräten und lassen sich über natürliche Sprache bedienen. Nicht die Mitarbeitenden müssen sich der Technologie anpassen, sondern die Technologie passt sich ihnen an. Steuerungsprozesse, die auf formale Beschaffungs- und Einführungszyklen ausgelegt sind, können mit dieser Dynamik nur schwer Schritt halten.

Damit können genau jene Kontrollpunkte umgangen werden, an denen Organisationen bisher Risiken erkannt, Entscheide getroffen und dokumentiert haben. Die Folge ist keine einzelne Schwachstelle, sondern ein Transparenzverlust: Die Organisation kann ihren eigenen Datenfluss nicht mehr vollständig beschreiben. Das betrifft unmittelbar auch den Datenschutz[2], weil Personendaten in Bearbeitungen gelangen können, die weder dokumentiert noch vertraglich geregelt sind.

Ein strukturiertes Inventar der eingesetzten KI-Anwendungsfälle und KI-Funktionen ist deshalb ein praktischer Ausgangspunkt jeder KI-Governance. Es unterstützt zugleich die Umsetzung eines Managementsystems nach ISO/IEC 42001, wie Kapitel 6 zeigt. Auch die Aufsicht setzt hier an: Die FINMA hat bei Instituten mit vielen oder wesentlichen KI-Anwendungen beurteilt, ob eine KI-Governance mit einem zentral geführten Inventar samt Risikoklassifizierung und daraus folgenden Massnahmen besteht. Sie hielt zugleich fest, dass die Vollständigkeit solcher Inventare eine Herausforderung darstellt, weil KI-Entwicklung und -Nutzung breit gestreut und generative Anwendungen für jedermann zugänglich sind.[3]

1.1 Drei Entstehungspfade von Schatten-KI

In der Praxis entsteht Schatten-KI über drei unterschiedliche Wege. Sie werden häufig vermischt, erfordern aber unterschiedliche Steuerungsantworten.

- **Individuelle Nutzung:** Mitarbeitende nutzen frei verfügbare Dienste für ihre tägliche Arbeit. Das Motiv ist nach der Erfahrung aus der Praxis meist Produktivität, nicht Regelverstoss, besonders wenn ein zugelassener Weg fehlt oder er zu umständlich ist.
- **Eingebettete Funktionen:** Anbieter bereits freigegebener Anwendungen aktivieren KI-Funktionen im Rahmen eines Releases, oder sie werden in der Konfiguration ohne Beurteilung eingeschaltet. Der Dienstleister und die Anwendung sind erfasst, die Funktion und allfällige weitere Bearbeiter nicht. Aus Governance-Sicht ist dies der anspruchsvollste Pfad, weil die Aktivierung meist vertraglich gedeckt ist, ein eigener Entscheid der Organisation aber häufig ausbleibt. Aktiviert der Anbieter die Funktion, haben Regeln für Mitarbeitende keinen Einfluss, weil es in der Organisation keine Handlung

[2] Datenschutz meint hier durchgehend den Schutz von Personendaten im Sinne des Datenschutzrechts. Informationssicherheit umfasst den Schutz aller schutzwürdigen Informationen, unabhängig davon, ob es sich um Personendaten handelt.
[3] vgl. FINMA-Aufsichtsmitteilung 08/2024 vom 18. Dezember 2024 zu Governance und Risikomanagement beim Einsatz von künstlicher Intelligenz.

gibt, an die sie anknüpfen könnten. Erschwerend kommt hinzu, dass der Anbieter häufig eine Black Box bleibt. Die FINMA hielt in derselben Aufsichtsmitteilung fest, dass Beaufsichtigte bei eingekauften Anwendungen teilweise Schwierigkeiten hatten, überhaupt zu eruieren, ob KI enthalten ist und welche Daten und Methoden verwendet werden.

- **Fachbereichseigene Automatisierung:** Abläufe, Auswertungen oder Agenten entstehen im Fachbereich und werden produktiv genutzt. Inventarisierung, Risikobeurteilung und Change-Prozess finden nicht statt, Informationssicherheit und Datenschutz bleiben aussen vor. Der Grund ist selten Umgehung: Was im Fachbereich als Arbeitserleichterung beginnt, wird nicht als Vorhaben wahrgenommen und durchläuft deshalb auch keines der dafür vorgesehenen Verfahren.

Allen drei Pfaden ist eines gemeinsam: Informationen und Prozesse geraten in eine Bearbeitung, die von der Organisation nie beurteilt und nie bewusst entschieden wurde. Schatten-KI ist damit in erster Linie ein Befund zur Steuerung und erst in zweiter Linie ein Befund zur KI. Sie zeigt eine Lücke in der Werttransparenz, also im Wissen, welche Informationen wohin gelangen und wo sie bearbeitet werden. Wo Mitarbeitende ausweichen, zeigt sie zudem, dass der freigegebene Weg den legitimen Bedarf nicht deckt.

Die folgenden Begriffe werden im gesamten Papier einheitlich verwendet.

Begriffe

Schatten-KI: Nutzung von KI, die stattfindet, ohne von der Organisation bewusst entschieden, dokumentiert oder verantwortet zu sein. Sie entsteht durch individuelle Nutzung, durch vom Anbieter aktivierte Funktionen oder durch fachbereichseigene Automatisierung (Abschnitt 1.1).

KI-Anwendungsfall: Ein bestimmter Zweck, für den KI in einem Prozess oder einer Tätigkeit eingesetzt wird. Er ist die Einheit für Inventar, Einstufung, Freigabe und Ergebnisverantwortung.

KI-Funktion: Eine KI-Fähigkeit innerhalb einer bestehenden Anwendung oder eines Cloud-Dienstes, die der Anbieter bereitstellt oder aktiviert.

Integrations- und Autonomiestufe: Die Einordnung eines Anwendungsfalls nach Integrationsgrad und Handlungsspielraum der KI (Kapitel 3). Sie beschreibt weder den Reifegrad der Organisation noch die rechtliche Einordnung.

Kanal und freigegebener Weg: Ein Kanal ist eine Kategorie der Bearbeitung, etwa ein öffentlich verfügbarer Dienst ohne Vertrag, ein Unternehmensmandant mit geregelter Auftragsbearbeitung oder eine intern betriebene Lösung. Der freigegebene Weg umfasst die konkreten Dienste, die für bestimmte Datenklassen zugelassen sind. «Freigegeben» und «zugelassen» werden gleichbedeutend verwendet.

Ergebnisverantwortung: Die Verantwortung für Nutzen, fachliche Verwendung und Prozessrisiko eines Anwendungsfalls, im RACI-Prinzip «Accountable». Sie liegt beim fachlichen Leistungsverantwortlichen (Business Owner). Bei individueller Nutzung auf Stufe 1 verantwortet die nutzende Person die Verwendung des Ergebnisses.

Werttransparenz: Das Wissen, welche Informationen welchen Wert haben, wer für sie verantwortlich ist und wo sie liegen. Bei KI gehört dazu, in welche Bearbeitungen Informationen gelangen.

Schutzbedarf: Das Schutzniveau, das eine Anwendung oder ein Anwendungsfall je Schutzziel (Vertraulichkeit, Integrität, Verfügbarkeit, Nachvollziehbarkeit) benötigt. Im Unterschied zur Datenklassifizierung betrifft er nicht nur die Vertraulichkeit.

Datenklassifizierung: Die Einstufung von Informationen in Datenklassen (z. B. intern, vertraulich) mit Vorgaben für Umgang, Kennzeichnung, Weitergabe und Ablage.

«Anbieter» und «Dienstleister» werden gleichbedeutend verwendet und schliessen interne Leistungserbringer ein.

1.2 Ziel der KI-Governance: kontrollierte Entwicklungsfähigkeit

Aus diesem Befund folgt, wozu Steuerung dient. KI-Governance soll die Risiken des KI-Einsatzes auf ein tragbares Mass begrenzen und zugleich die Voraussetzungen dafür schaffen, dass KI bewusst eingesetzt werden kann. Beides gehört zusammen. Eine Steuerung, die nur begrenzt, verliert die Nutzung aus dem Blick, statt sie zu beherrschen. Eine Steuerung, die nur ermöglicht, gibt die Transparenz auf, die sie für jede spätere Beurteilung benötigt.

Das Ziel lässt sich deshalb als kontrollierte Entwicklungsfähigkeit beschreiben. Die Organisation weiss, wo und wofür KI eingesetzt wird, wer die Ergebnisverantwortung trägt und unter welchen Voraussetzungen sich die Nutzung weiterentwickeln darf. Sie kann den nächsten Schritt bewusst gehen, statt ihn unbemerkt zu vollziehen. Kapitel 3 beschreibt diese Schritte als Integrations- und Autonomiestufen.

Diese Zielsetzung hat eine praktische Konsequenz für den Aufbau. Eine Governance, die den legitimen Bedarf des Geschäfts nicht abdeckt oder zu langsam entscheidet, erzeugt genau jene Schatten-KI, die sie verhindern soll. Ihre Wirksamkeit bemisst sich deshalb nicht am Umfang der Regelwerke, sondern zuerst daran, ob Entscheide zeitnah, nachvollziehbar und auf belastbarer Grundlage getroffen werden. Wie sich die Wirksamkeit nachweisen lässt, zeigt Kapitel 7.

2 Rechtlicher Rahmen im Überblick

Ein verbreiteter Irrtum lautet, der Einsatz von KI sei in der Schweiz noch ungeregelt und die rechtliche Klärung könne warten, bis ein KI-Gesetz vorliegt. Richtig ist, dass bislang keine übergreifende KI-spezifische Gesetzgebung besteht. Ebenso richtig ist aber, dass der Einsatz von KI bereits heute von mehreren bestehenden Rechtsgebieten erfasst wird. Die folgende Übersicht bleibt bewusst auf der Ebene der Rechtsgebiete und ersetzt keine rechtliche Beurteilung im Einzelfall.[4] Je nach Trägerschaft und Branche kommen weitere Vorgaben hinzu, bei Bundesbehörden etwa die Informationssicherheitsgesetzgebung, bei Kantonen und Gemeinden die jeweiligen kantonalen Erlasse.

Rechtsgebiet	Worum es geht	Bezug zur KI-Nutzung
Datenschutz *Bearbeitungsgrundsätze und Datensicherheit*	Zweckbindung, Verhältnismässigkeit und Erkennbarkeit der Bearbeitung. Angemessene technische und organisatorische Schutzmassnahmen.	Die Eingabe von Personendaten in ein Modell ist eine Bearbeitung im Sinne des Datenschutzrechts. Zweck, Schutzniveau und Zulässigkeit sind vor der Nutzung zu bestimmen, nicht danach.
Datenschutz *Auftragsbearbeitung und Bekanntgabe ins Ausland*	Vertragliche Regelung der Bearbeitung durch Dritte. Genehmigung beim Beizug weiterer Auftragsbearbeiter. Anforderungen an die Bekanntgabe ins Ausland.	Aktiviert ein Bestandsanbieter eine KI-Funktion und zieht dafür weitere Bearbeiter bei, greifen die vereinbarten Informations- und Genehmigungsmechanismen. Verarbeitungsort und Verwendungszwecke sind neu zu beurteilen.

[4] Massgebend sind insbesondere das Datenschutzgesetz (DSG, SR 235.1) mit der zugehörigen Verordnung, Art. 328b OR und Art. 26 ArGV 3, die Geheimhaltungsbestimmungen des Strafgesetzbuches und der Spezialgesetzgebung sowie sektorielle Vorgaben.

Rechtsgebiet	Worum es geht	Bezug zur KI-Nutzung
Datenschutz *Automatisierte Entscheide, Folgenabschätzung und Meldepflicht*	Besondere Anforderungen bei Entscheiden, die ohne menschliche Beteiligung ergehen und erheblich wirken. Folgenabschätzung bei potenziell hohem Risiko. Meldepflicht bei Verletzungen der Datensicherheit mit voraussichtlich hohem Risiko.	Relevant, wenn der Entscheid ohne menschliche Beteiligung ergeht und die gesetzlichen Voraussetzungen hinsichtlich seiner Wirkung erfüllt sind. Gemeint ist nicht jede computergestützte Bearbeitung. Bei entscheidungsunterstützendem Einsatz bleibt die tatsächliche Qualität der menschlichen Überprüfung governancerelevant.
Arbeitsrecht *Schutz der Persönlichkeit am Arbeitsplatz*	Bearbeitung von Arbeitnehmerdaten nur, soweit sie die Eignung für das Arbeitsverhältnis betreffen oder zur Durchführung des Arbeitsvertrags erforderlich sind.	Setzt der personenbezogenen Verhaltenskontrolle enge Grenzen. Technische Kontrollen, die Daten und Kanäle steuern, sind davon zu unterscheiden, soweit sie nicht der Verhaltensüberwachung dienen.
Geheimhaltungspflichten *Amtsgeheimnis, Berufsgeheimnisse, Geschäftsgeheimnisse*	Pflichten, die unabhängig vom Datenschutzrecht bestehen und teilweise strafbewehrt sind. Dazu kommen vertragliche Vertraulichkeitszusagen.	Die Eingabe besonders schützenswerter Personendaten oder geschützter Geschäftsinformationen wie Forschungs- und Entwicklungsdaten in einen nicht freigegebenen Dienst kann über das Datenschutzrecht hinaus relevant werden.
Aufsichtsrecht *Sektorielle Vorgaben und Aufsichtspraxis*	Erwartungen an Governance, Inventar und Risikoklassifizierung, Datenqualität, Erklärbarkeit, Modellüberwachung und Drittparteienabhängigkeit.	Aufsichtsbehörden leiten ihre Erwartungen bereits heute aus bestehendem Recht ab, ohne ein KI-Gesetz abzuwarten.

Tabelle 1: Rechtsgebiete mit Bezug zur KI-Nutzung in der Schweiz (nicht abschliessend)

2.1 Einordnung des EU-Rechts

Das EU-Recht ist in der KI-Regulierung deutlich weiter fortgeschritten.

> **Stufenplan der EU-KI-Verordnung (Stand Oktober 2026)**
>
> Seit Februar 2025 gelten die bisherigen Verbote nach Art. 5 und die Vorgaben zur KI-Kompetenz nach Art. 4. Der «Digital Omnibus on AI» (Verordnung (EU) 2026/1744) hat Art. 4 neu gefasst: Verlangt sind Massnahmen zur Förderung der KI-Kompetenz, ein bestimmtes Niveau muss nicht garantiert werden. Seit August 2025 gelten die Pflichten für Modelle mit allgemeinem Verwendungszweck, seit August 2026 die Transparenzpflichten nach Art. 50. Für die maschinenlesbare Kennzeichnung nach Art. 50 Abs. 2 gilt bei vor dem 2. August 2026 in Verkehr gebrachten Systemen zur Erzeugung synthetischer Audio-, Bild-, Video- oder Textinhalte eine Übergangsfrist bis 2. Dezember 2026. Ab diesem Datum gelten zudem neue Verbote nach Art. 5. Die Pflichten für Hochrisiko-Systeme (Kapitel III, Abschnitte 1 bis 3, mit Ausnahme von Art. 6 Abs. 5) wurden verschoben: für eigenständige Systeme nach Art. 6 Abs. 2 und Anhang III auf den 2. Dezember 2027, für KI in Produkten nach Art. 6 Abs. 1 und Anhang I auf den 2. August 2028. Diese Verschiebung ist kein Aufschub der Governance-Arbeit, sondern das Zeitfenster, in dem sie entstehen kann.

Für Schweizer Organisationen ist das EU-Recht in zweifacher Hinsicht relevant. Unmittelbar, weil die KI-Verordnung sie erfassen kann, etwa wenn sie KI-Systeme auf dem Markt der Union in Verkehr bringen oder wenn die von einem KI-System erzeugten Ergebnisse dort verwendet werden. Und als Referenz, weil sie die

Richtung vorgibt, an der sich Erwartungen von Kunden, Partnern und Aufsicht bereits heute orientieren. In der Schweiz wird bis Ende 2026 eine Vernehmlassungsvorlage zur Umsetzung der KI-Konvention des Europarats erarbeitet, mit Schwerpunkten bei Transparenz, Datenschutz, Nichtdiskriminierung und Aufsicht.

3 Von assistiver Nutzung zu agentischen Systemen

Wie geht eine Organisation mit einer Nutzung um, die bereits stattfindet? Ein Verbot greift zu kurz, weil es keinen der drei Entstehungspfade zuverlässig erreicht: Bei individueller Nutzung verlagert es diese auf nicht kontrollierte Kanäle, bei fachbereichseigener Automatisierung greift es erst, wenn sie bekannt ist, und bei Funktionen, die Anbieter selbst aktivieren, geht es ins Leere. Zielführender ist es, den Umgang mit KI als Entwicklungspfad zu verstehen und je Stufe zu klären, welche Voraussetzungen erfüllt sein müssen.

Der Umgang mit KI entwickelt sich in Organisationen typischerweise entlang von vier Stufen. Sie beschreiben eine zunehmende Integration und Autonomie: von der Bearbeitung bereitgestellter Informationen über assistive Funktionen, die Anbieter in bestehende Anwendungen einbetten, und den Zugriff auf organisationseigene Daten und Prozesse bis zur eigenständigen Ausführung von Handlungen. Verkürzt lautet die Abfolge: assistieren, einbetten, integrieren, handeln. Eingestuft wird der einzelne Anwendungsfall, nicht die Organisation. Eine Organisation nutzt in der Regel mehrere Stufen nebeneinander. Im Modell beschreibt die Stufe den Integrationsgrad und den Handlungsspielraum eines Anwendungsfalls. Sie ist weder ein Reifegrad der Organisation noch eine rechtliche Einordnung. Die rechtlichen Anforderungen richten sich nach den in Kapitel 2 beschriebenen Kriterien und können auf jeder Stufe unterschiedlich weit reichen.

Mit jeder Stufe verändert sich auch die Art des Risikos. Auf den unteren Stufen bearbeitet KI Informationen, ohne selbst Handlungen in Geschäftsprozessen auszulösen. Im Vordergrund stehen dort Vertraulichkeit, Datenschutz und Geheimhaltungspflichten. Mit zunehmender Integration kommen Risiken für Integrität, Nachvollziehbarkeit und Verfügbarkeit hinzu. Die Risiken der tieferen Stufen verschwinden dabei nicht. Sie akkumulieren, denn ein Agent mit weitreichendem Datenzugriff trägt das Vertraulichkeitsrisiko in verschärfter Form weiter. Die Stufe bestimmt damit vor allem, welche Art von Steuerung erforderlich ist. Wie weit diese gehen muss, hängt zusätzlich vom Schutzbedarf des Anwendungsfalls, von der Kritikalität des Prozesses und von den möglichen Auswirkungen auf betroffene Personen ab. Ein sensibler Anwendungsfall auf Stufe 1 kann deshalb anspruchsvoller sein als ein unkritischer Agent.

Für die Steuerung sind dabei drei Beobachtungen wesentlich:

- Die Anforderungen sind kumulativ. Ein Anwendungsfall auf Stufe 3 muss auch die Grundanforderungen der tieferen Stufen erfüllen, etwa die Nutzungsregeln und den freigegebenen Weg. Wer Stufe 3 nutzt, betreibt zudem meist auch die Stufen 1 und 2 weiter.
- Stufe 2 nehmen viele Organisationen unbemerkt ein, weil die Nutzung nicht durch einen Entscheid entsteht, sondern durch ein Update des Anbieters oder eine ungeprüft aktivierte Funktion. Für den Übergang in Stufe 3 oder 4 ist die Wirkung massgebend, nicht der Bereitstellungsweg: Sobald eine eingebettete Funktion eigenständig auf Datenbestände zugreift oder Prozesse beeinflusst, gehört sie in Stufe 3 oder 4.
- Der Übergang von Stufe 2 zu Stufe 3 ist ein Wechsel in der Risikologik. Ab hier werden Identitäts- und Zugriffsmanagement (IAM), Change-Management und das interne Kontrollsystem tragend.

Tabelle 2 fasst die vier Stufen mit ihrer typischen Nutzung, den dominierenden Risiken und den Governance-Voraussetzungen zusammen.

Stufe	Typische Nutzung	Dominierende Risiken	Governance-Voraussetzungen
1 – Assistive Nutzung *KI als persönliches Werkzeug*	Formulieren, Zusammenfassen, Recherchieren über Chat-Oberflächen. Das Ergebnis wird individuell verwendet	Vertraulichkeit, Datenschutz und Geheimhaltungspflichten. Unkontrollierter Abfluss von Informationen. Qualität ungeprüfter Ergebnisse	Nutzungsregeln entlang der Datenklassifizierung, freigegebener Kanal, Awareness. Die Verantwortung für die Verwendung des Ergebnisses bleibt bei der Person
2 – Eingebettete assistive Funktionen *KI als Produktmerkmal*	Assistive KI-Funktionen in bereits genutzten Fachanwendungen und Cloud-Diensten, häufig durch den Anbieter aktiviert. Die Nutzung bleibt unter menschlicher Kontrolle	Bearbeitung ohne bewussten Entscheid. Fehlende Transparenz über Bearbeitungsort, Trainingsnutzung und weitere Bearbeiter. Manipulation über verarbeitete Inhalte	Drittparteiensteuerung: Inventar auf Funktionsebene, vertragliche Informationspflicht bei Funktionsänderungen, Prüfung der Mandantenkonfiguration
3 – Prozessintegrierte KI *KI als System- und Prozessbestandteil*	KI mit Zugriff auf organisationseigene Daten. Ergebnisse fliessen in Geschäftsprozesse ein	Verschiebung zu Integrität und Nachvollziehbarkeit. Datenqualität. Zu breite Datenzugriffe über vererbte Berechtigungen. Manipulation von Ein- und Ausgaben. Aggregation breiter Datenbestände. Verfügbarkeit und Abhängigkeit vom Anbieter	Schutzbedarfs- und Risikoanalyse vor Freigabe, IAM-Konzept für den Datenzugriff, Qualitätskriterien und Validierung der Ein- und Ausgaben, Change- und Testverfahren, benannte fachliche Verantwortung, Ausweich- oder Notbetrieb bei kritischen Prozessen
4 – Agentische Systeme *KI als handelnder Akteur im Prozess*	Eigenständige Ausführung von Handlungen in angebundenen Systemen, mit interner oder externer Wirkung	Integrität, Nachvollziehbarkeit und Verfügbarkeit. Fehlerhafte Aktionen skalieren. Abhängigkeit im kritischen Prozess	Eindeutige technische Identität je produktivem Agenten, Least Privilege und Rezertifizierung, zugelassene Tools und Funktionen, Aktions- und Transaktionsgrenzen, menschliche Freigabe bei irreversiblen oder folgenreichen Aktionen, Protokollierung, Abschalt- und Rückabwicklungsfähigkeit, Einbezug ins Business Continuity Management (BCM)

Tabelle 2: Integrations- und Autonomiestufen der KI-Nutzung und ihre Governance-Voraussetzungen

Das Modell beschreibt keine Entwicklungspflicht. Nicht jede Organisation muss Stufe 4 erreichen. Wer eine Stufe erreicht, ohne deren Voraussetzungen erfüllt zu haben, hat lediglich schneller ein grösseres Problem.

Eine Unterscheidung, die in der Praxis regelmässig fehlt: Die vier Stufen beschreiben, wie tief KI eingebunden ist und wie viel Handlungsspielraum sie erhält. Sie sagen nichts darüber aus, wie gut die Organisation diesen Einsatz steuert. Zwei Organisationen können beide agentische Systeme betreiben: Die eine nutzt Identitäten, Freigabeschwellen, Protokollierung und Rückabwicklung, die andere hat die Systeme lediglich produktiv geschaltet.

KI-spezifische Angriffsflächen gewinnen mit der Integration an Gewicht

Mit zunehmender Integration verändert sich nicht nur die Governance-, sondern auch die Cybersicherheits-Perspektive. Zu den KI-spezifischen Angriffsflächen gehören manipulierte Eingaben, die unerwünschte Offenlegung von Informationen, die unsichere Weiterverarbeitung von Modell-Ausgaben, Risiken in der Modell- und Lieferkette sowie eine zu weitreichende autonome Handlungsfähigkeit.[5] Viele davon bestehen schon auf den unteren Stufen, etwa wenn ein eingebetteter Assistent manipulierte Inhalte aus E-Mails oder Dokumenten verarbeitet. Mit der Integration wächst jedoch ihr Schadenspotenzial: Was auf Stufe 1 zu einer Offenlegung führt, kann auf Stufe 4 eine Handlung auslösen.

Besonders deutlich wird das, wo KI in kritische Prozesse oder in Infrastrukturen mit Versorgungsrelevanz einfliesst und Integritäts- oder Verfügbarkeitsprobleme unmittelbar auf die Leistungserbringung durchschlagen. KI-Governance und Cyber- beziehungsweise Informationssicherheits-Governance lassen sich mit zunehmendem Integrationsgrad immer weniger getrennt betrachten.

Ein hoher Integrationsgrad ist kein hoher Reifegrad. Reife zeigt sich darin, ob die Governance mit dem Integrationsgrad Schritt hält.

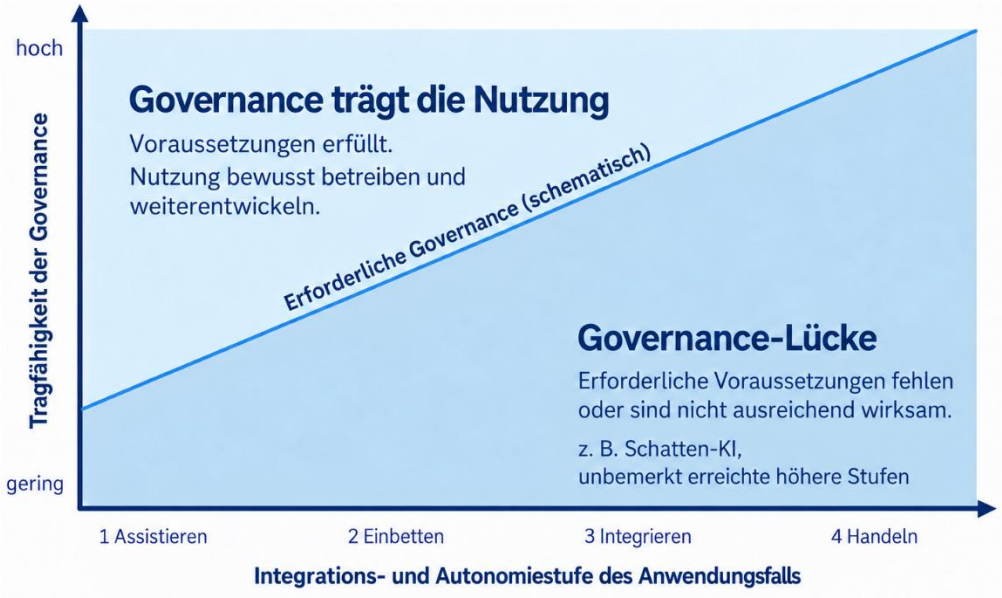


Abbildung 1: Integrations- und Autonomiestufe des Anwendungsfalls und erforderliche Governance (schematisch, eigene Darstellung)

4 Elemente wirksamer Steuerung

4.1 Regeln, die im Alltag tragen

Mahnungen zur Vorsicht genügen nicht. Wirksame Nutzungsregeln sind Entscheidungshilfen. Sie docken an die bestehende Schutzbedarfs- und Datenklassifizierungslogik an und beantworten drei Fragen. Welche Datenklasse darf in welchem Kanal bearbeitet werden? Wer entscheidet bei Grenzfällen? Und wie lange dauert dieser Entscheid?

[5] Die genannten Angriffsflächen orientieren sich an den Risikokategorien der OWASP Top 10 for LLM Applications (Fassung 2025) des OWASP GenAI Security Project, darunter Prompt Injection, Sensitive Information Disclosure, Supply Chain, Improper Output Handling und Excessive Agency.

Die dritte Frage wird regelmässig unterschätzt. Benötigt die Freigabe eines einfachen assistiven Anwendungsfalls mehrere Wochen, ist die Regel faktisch wirkungslos, unabhängig davon, wie sorgfältig sie formuliert wurde. Für prozessintegrierte und agentische Anwendungsfälle ist eine gründlichere Beurteilung angemessen, sofern Kriterien, Weg und Dauer vorab bekannt sind. Eine Entscheidgeschwindigkeit, die zu Risiko und Komplexität passt, ist damit selbst ein Sicherheitsmerkmal.

Regeln sollten technologieoffen formuliert sein. Listen einzelner zugelassener KI-Werkzeuge veralten schneller, als sie gepflegt werden können, und genügen deshalb allein nicht als Regel. Tragfähiger sind Kriterien: welche Datenklasse, welcher Schutzbedarf, welche vertragliche Grundlage, welcher Bearbeitungsort. Kanäle sind dabei Kategorien der Bearbeitung (vgl. Begriffe in Kapitel 1). Ihnen werden die konkret freigegebenen Dienste zugeordnet, die für die Praxis wichtig bleiben. Der zugelassene Kanal muss zudem die tatsächlichen Anwendungsfälle des Geschäfts abdecken, sonst entstehen unsichtbare Workarounds. Awareness ergänzt die Regeln und wirkt dort am besten, wo sie an konkreten Beispielen aus dem Arbeitsalltag ansetzt.

Technische Kontrollen gehören dazu. Risikobasierte Massnahmen wie Zugriffssteuerung, Mandantenkonfigurationen, Schutz vor Datenabfluss oder die Steuerung der zugelassenen Kanäle ergänzen die organisatorischen Regeln wirksam und setzen präventiv an. Sie steuern Daten und Kanäle, nicht Personen. Eine personenbezogene Auswertung des individuellen Nutzungsverhaltens zum Zweck der Verhaltenskontrolle ist dagegen kritisch. Sie kann Ausweichverhalten fördern und stösst an arbeits- und datenschutzrechtliche Grenzen.

4.2 Was über Menschen und was über Verträge gesteuert wird

Nicht jedes Risiko lässt sich mit denselben Mitteln steuern. Regeln und Awareness wirken dort, wo Menschen handeln. Sie adressieren den ersten und den dritten Entstehungspfad. Gegen den zweiten Pfad helfen sie nur begrenzt: Sie können verlangen, dass neue Funktionen vor dem Einschalten beurteilt werden, und regeln, wie sie genutzt werden. Sie erreichen weder die Aktivierung durch den Anbieter noch den Beizug weiterer Bearbeiter (Abschnitt 1.1).

Dieser Pfad gehört vor allem zur Lieferantensteuerung. Zieht ein Anbieter für eine neue KI-Funktion weitere Dienstleister bei, greifen die vertraglich und datenschutzrechtlich vereinbarten Informations- und Genehmigungsmechanismen. Betroffen sind dabei nicht nur Personendaten, sondern ebenso Geschäftsgeheimnisse und vertragliche Vertraulichkeitszusagen. Eine Produktneuerung des Anbieters entbindet die Auftraggeberin also nicht von ihrer Entscheidung, sondern verlangt, dass diese vorbereitet ist. Die Verantwortung bleibt intern, wie das Whitepaper IT-Outsourcing & Governance dieser Reihe ausführt.

Vier Instrumente sind dafür zentral: vertraglich verankerte Informationspflichten bei wesentlichen Funktionsänderungen, Transparenz über weitere Bearbeiter, Klarheit über Bearbeitungsort und Trainingsnutzung der Daten sowie die regelmässige Überprüfung der Mandantenkonfiguration. Ob sich eine Funktion zentral deaktivieren lässt, unterscheidet sich dabei je nach Anbieter und ist vorab zu klären. Wo die Konfiguration den Bearbeitungsumfang tatsächlich steuert, ist sie ein wirksames Instrument. Wo sie es nicht tut, bleibt die vertragliche Ebene. Reichen auch technische und vertragliche Bedingungen nicht aus, bleiben Einschränkung, Anbieterwechsel oder der Verzicht auf die Funktion.

Bei wesentlichen Anwendungsfällen umfasst die Anbietersteuerung auch die Verfügbarkeit. Zu beurteilen sind Abhängigkeiten von Modell-, Cloud- und Plattformanbietern, auch wenn mehrere Verträge auf demselben Modell oder derselben Infrastruktur beruhen. Ob ein Modell ersetzt werden kann, ist keine reine Vertragsfrage, sondern hängt von der Gestaltung des Anwendungsfalls ab. Welche Ausweich- oder Notbetriebslösung trägt, ist Gegenstand des Business Continuity Managements.

4.3 Verantwortung und Organisation

Verantwortlichkeiten lassen sich entlang des Lebenszyklus eines Anwendungsfalls beschreiben: Antrag, Einstufung (Schutzbedarf, Datenklassifizierung, Risiko), Freigabe, Betrieb, Überwachung, Änderung und Ausserbetriebnahme. Geeignet ist eine Zuordnung nach dem RACI-Prinzip (Responsible, Accountable, Consulted, Informed), das je Aktivität zwischen Durchführung, Ergebnisverantwortung, Beizug und Information unterscheidet. Entscheidend ist, dass sie an diesen Aktivitäten ansetzt und nicht pauschal an Organisationseinheiten.

Die häufigste Schwäche in der Praxis ist nicht eine fehlende Matrix, sondern eine Ergebnisverantwortung, die faktisch niemand ausfüllt. Informationssicherheit, Datenschutz und IT werden als ergebnisverantwortlich (Accountable) eingetragen, obwohl sie den fachlichen Nutzen nicht verantworten können. Die Ergebnisverantwortung für einen KI-Anwendungsfall gehört zum fachlichen Leistungsverantwortlichen (Business Owner), bei dem auch das Prozessrisiko liegt. Er verantwortet Nutzen und fachliche Verwendung. Sollen organisationseigene Informationen genutzt werden, entscheidet der Informationseigner (Data Owner) über den Zugriff. Die Kontroll- und Unterstützungsfunktionen beraten, prüfen und eskalieren. Die Ergebnisverantwortung übernehmen sie nicht, auch dort nicht, wo sie, etwa für Datenschutz oder Recht, verbindlich über die Zulässigkeit entscheiden.

Für die laufende Steuerung braucht es zudem einen klaren Entscheidweg mit definierter Freigabekompetenz. Je nach Organisationsgrösse kann diese bei einer benannten Funktion, bei einem bestehenden Security- oder Risikogremium oder bei einem eigenen KI-Board liegen. Wo ein Gremium eingerichtet wird, sollte es aus bestehenden Rollen zusammengesetzt sein und kurze Wege ermöglichen. Sinnvoll vertreten sind die fachlichen Leistungsverantwortlichen, Informationssicherheit, Datenschutz, IT und Architektur sowie fallweise Recht und Beschaffung. Wichtiger als die Form ist, dass entschieden statt nur beraten wird und dass die Entscheide dokumentiert sind und dort ankommen, wo sie umgesetzt und nachverfolgt werden. Dabei sind zwei Entscheide zu unterscheiden. Die zuständige Freigabestelle beurteilt, ob die festgelegten Voraussetzungen erfüllt sind. Verbleibende Risiken werden von der dafür zuständigen und ausreichend bevollmächtigten Stelle dokumentiert akzeptiert oder an die Leitung eskaliert. Eine interne Risikoakzeptanz ersetzt keine gesetzliche Zulässigkeit.

5 KI-Governance aufbauen: Reihenfolge und Prioritäten

Die Reihenfolge prägt den Erfolg. Viele Organisationen beginnen mit einer Weisung. Das ist naheliegend, aber selten wirksam, weil eine Regel ohne Kenntnis der tatsächlichen Nutzung nicht zielgerichtet sein kann. Der Aufbau erfolgt deshalb in sechs Schritten. Sie beschreiben Schwerpunkte, keine strikte Abfolge. Erkennbare Risiken werden sofort begrenzt, ohne den Abschluss der übrigen Schritte abzuwarten.

1. Transparenz schaffen. Ein Inventar der genutzten Anwendungsfälle und der KI-Funktionen in Bestandsanwendungen aufbauen, ausdrücklich einschliesslich der bestehenden Schatten-KI. Quellen sind Beschaffungsunterlagen, Release-Informationen der Anbieter und eine strukturierte Selbstdeklaration der Fachbereiche. Die Erstdeklaration bestehender Nutzung sollte ausdrücklich sanktionsfrei sein, sonst bleibt gerade die Schatten-KI verborgen. Ohne diesen Schritt bleibt alles Weitere Vermutung.

2. Verantwortung festlegen. Die Ergebnisverantwortung je Anwendungsfall benennen, eine koordinierende Funktion bestimmen und den Entscheidweg mit Freigabekompetenz festlegen. Dazu gehört ein niederschwelliger Meldeweg, über den Mitarbeitende Beobachtungen und Bedenken einbringen können. Wer auf Überwachung verzichtet, ist auf diese Hinweise angewiesen.

3. Regeln entlang der Datenklassifizierung formulieren. Festlegen, welche Datenklasse in welchem Kanal zulässig ist, wer Grenzfälle entscheidet und welche Nutzung ausgeschlossen bleibt.

4. Einen zugelassenen Weg bereitstellen. Regeln ohne verfügbare Alternative erzeugen Ausweichnutzung. Der freigegebene Weg muss den legitimen Bedarf tatsächlich abdecken und bekannt sein. Regeln und freigegebener Weg treten deshalb gemeinsam in Kraft. Eine geschützte Testumgebung erlaubt es, neue Werkzeuge zu erproben, ohne dass jemand ausweichen muss.

5. Freigabe- und Risikoprozess verankern. Neue Anwendungsfälle einstufen. Die Einstufung verbindet Integrationsstufe, Schutzbedarf des Anwendungsfalls, Kritikalität des Prozesses und rechtliche Einordnung (Kapitel 2). Sie bestimmt Prüftiefe, die erforderliche Folgenabschätzung (Kapitel 6) und die Freigabeebene. Für wesentliche Anwendungsfälle Qualitätskriterien und Grenzen der Verwendung festlegen. Den Entscheid dokumentieren, einschliesslich der Stellen, an denen ein Mensch zwingend eingreift, und verbleibende Risiken ausdrücklich akzeptieren lassen (Abschnitt 4.3). Der Prozess ist an Beschaffung und Change-Management anzubinden.

6. Drittparteien und technische Kontrollen einbeziehen. Vertragliche Anforderungen an Anbieter ergänzen und risikobasierte technische Kontrollen definieren, die Daten und Kanäle steuern.

Der Nachweis der Wirksamkeit ist kein siebter Schritt, sondern begleitet alle sechs. Kapitel 7 systematisiert ihn.

Nicht alle Elemente haben dieselbe Priorität:

Priorität	Elemente
Basis	Inventar der Anwendungsfälle und der KI-Funktionen in Bestandsanwendungen, benannte Ergebnisverantwortung, Nutzungsregeln entlang der Datenklassifizierung, ein zugelassener Weg, der geregelte Beizug von Informationssicherheit und Datenschutz bei der Beurteilung neuer Anwendungsfälle sowie ein niederschwelliger Meldeweg. Diese Elemente sind auch für kleine Organisationen mit begrenzten Mitteln realistisch.
Kurzfristig	Freigabe- und Risikoprozess, vertragliche Anforderungen an Anbieter samt Transparenz über Funktionsänderungen, Anbindung an Beschaffung und Change-Management, zielgruppengerechtes Awareness-Konzept.
Risikobasiert	Weitergehende technische Kontrollen, spezifische Kontrollen für agentische Systeme, Einbezug ins Business Continuity Management, Ausweich- oder Notbetrieb und Exit bei kritischen Anbieterabhängigkeiten, formalisiertes Managementsystem nach ISO/IEC 42001.

Tabelle 3: Priorisierung der Elemente einer KI-Governance

Welche technischen Kontrollen erforderlich sind, hängt von Integrationsstufe, Schutzbedarf und Risikoexposition ab. Mandantenkonfigurationen und die Steuerung nicht freigegebener Kanäle können bereits auf den Stufen 1 und 2 relevant sein. Eigene technische Identitäten, Aktionsgrenzen, erweiterte Protokollierung und die Einbindung ins Business Continuity Management gewinnen vor allem auf den Stufen 3 und 4 an Bedeutung.

6 ISO/IEC 42001 als Steuerungsrahmen

Wer für den Aufbau einen etablierten Rahmen sucht, findet ihn in ISO/IEC 42001[6]. Die Norm beschreibt ein Managementsystem für künstliche Intelligenz (AIMS) mit Anforderungen an Kontext, Leitlinien der Leitung, Rollen, Risikobeurteilung, Folgenabschätzung, Betrieb und Verbesserung. An der Norm kann sich eine Organisation von Beginn an orientieren. Ein formalisiertes oder zertifiziertes AIMS ist dagegen ein risikobasierter Schritt (Tabelle 3). Die Norm folgt der harmonisierten Struktur der ISO-Managementsystemnormen, die auch ISO/IEC 27001 für Informationssicherheit und ISO 22301 für Business Continuity verwenden. Ein AIMS lässt sich deshalb in ein bestehendes Managementsystem für Informationssicherheit (ISMS) integrieren, statt als Parallelsystem geführt zu werden. Risikobeurteilung, internes Audit, Managementbewertung, Lieferantensteuerung und Vorfallbehandlung können mit gemeinsamen Abläufen und fachspezifischen Kriterien betrieben werden. Für Verfügbarkeit, Ausweich- oder Notbetrieb und Wiederanlauf besteht zudem eine Schnittstelle zum Business Continuity Management (vgl. Abbildung 2).



Abbildung 2: KI-Governance im integrierten Managementsystem (eigene Darstellung)

Operative Grundlage eines AIMS ist das Inventar aus Kapitel 5. Es erfasst neben den KI-Systemen auch die Anwendungsfälle und die KI-Funktionen in Bestandsanwendungen. Geltungsbereich, Risikobeurteilung, Folgenabschätzung und Überwachung im Betrieb lassen sich nur für erfasste Systeme bestimmen. Wer den Aufbau bei den Leitlinien statt beim Inventar beginnt, dokumentiert Absichten statt Zustände.

[6] ISO/IEC 42001:2023 (AI Management System). Ergänzend: ISO/IEC 23894 (Leitfaden KI-Risikomanagement) und ISO/IEC 42005:2025 (Folgenabschätzung von KI-Systemen).

Drei KI-spezifische Elemente erweitern die Perspektive eines klassischen ISMS besonders deutlich. Das erste ist die Folgenabschätzung für KI-Systeme. Sie knüpft am System an und nicht an der Bearbeitung von Personendaten. Damit erfasst sie auch Auswirkungen auf betroffene Personen und Gruppen, die ausserhalb des Anwendungsbereichs einer datenschutzrechtlichen Folgenabschätzung liegen. Die datenschutzrechtliche Folgenabschätzung bleibt daneben bestehen und beurteilt die Risiken einer Bearbeitung für die betroffenen Personen samt den Massnahmen zu deren Eindämmung. Für Betreiber bestimmter Hochrisiko-Systeme kann nach EU-Recht zudem eine Grundrechte-Folgenabschätzung erforderlich sein. Das zweite ist die Betrachtung des Lebenszyklus mit Themen wie Verzerrung, Erklärbarkeit und Veränderung des Modellverhaltens im Betrieb. Das dritte ist die fachliche Qualität: Ob ein KI-System für seinen Zweck geeignet ist und bleibt, lässt sich nicht aus seiner Sicherheit ableiten. Es braucht Qualitätskriterien, Tests und eine Überwachung der fachlichen Leistung im Betrieb. Ein klassisches ISMS adressiert diese Fragestellungen nicht in derselben KI-spezifischen Tiefe und Systematik.

Ebenso wichtig ist die Abgrenzung. Eine Zertifizierung nach ISO/IEC 42001 bestätigt die Konformität des Managementsystems innerhalb des definierten Geltungsbereichs. Sie belegt nicht, dass ein einzelner Anwendungsfall beherrscht ist, und ersetzt weder den anwendungsbezogenen Wirksamkeitsnachweis noch den Nachweis regulatorischer Konformität. Eine Konformitätsvermutung nach der EU-KI-Verordnung (Art. 40) begründet sie nicht: Diese entsteht nur durch harmonisierte Normen, deren Fundstellen im Amtsblatt der EU veröffentlicht sind. Der Nutzen der Norm liegt in der Struktur, die sie vorgibt, nicht im Nachweis, den sie liefert.

7 Von der Governance zur nachweisbaren Wirksamkeit

Eine Weisung, ein Inventar und ein Freigabeprozess sind noch keine wirksame Governance.

Die vorangehenden Kapitel beschreiben, was aufzubauen ist. Dieses Kapitel beantwortet die Anschlussfrage, woran sich erkennen lässt, ob es wirkt. Leitung, Aufsicht und Prüfer verlangen genau diesen Nachweis. Er fällt regelmässig schwer, weil KI-Nutzung dezentral, informell und schnell entsteht.

Design Effectiveness fragt, ob eine Kontrolle angemessen gestaltet ist. Operating Effectiveness fragt, ob sie über einen Zeitraum tatsächlich angewendet wurde und ob sich das belegen lässt. Die rechte Spalte der folgenden Übersicht beantwortet deshalb durchgehend dieselbe Frage: Welche Nachweise würde man sich zeigen lassen?

Design Effectiveness	Operating Effectiveness
Eine verbindliche Nutzungsregelung ist an die Datenklassifizierung gekoppelt und durch ein zielgruppengerechtes Awareness-Konzept begleitet.	Stichproben und Gespräche zeigen, dass die Zielgruppen die Regeln und den zugelassenen Weg kennen. Aggregierte, nicht personenbezogene Auswertungen der Kanal- und Datenkontrollen stützen, dass sie angewendet werden. Abweichungen werden behandelt und fliessen in die Awareness zurück.
Ein Inventar der KI-Anwendungsfälle und der KI-Funktionen in Bestandsanwendungen ist definiert, inklusive Verantwortlichkeit.	Das Inventar wird regelmässig mit Beschaffungen, Changes und Anbieter-Releases abgeglichen. Stichproben bestätigen die Aktualität, der Abgleich plausibilisiert die Vollständigkeit. Lücken werden zeitnah bereinigt.

Design Effectiveness	Operating Effectiveness
Freigabeprozess und Einstufungskriterien sind festgelegt, ebenso die daraus folgende Prüftiefe und Freigabeebene sowie die Zuständigkeit für die Akzeptanz verbleibender Risiken.	Stichproben von Freigaben zeigen, dass die Kriterien konsistent angewendet, Auflagen dokumentiert und nachverfolgt und Entscheide innerhalb angemessener Fristen getroffen werden. Verbleibende Risiken sind von der zuständigen Stelle akzeptiert.
Die Ergebnisverantwortung je Anwendungsfall ist beim fachlichen Leistungsverantwortlichen verankert.	Risikobeurteilungen, Freigaben, Ergebnisüberprüfungen und Eskalationen belegen, dass die benannte Verantwortung tatsächlich wahrgenommen wird.
Anforderungen an Anbieter sind vertraglich verankert, etwa zu Funktionsänderungen, weiteren Bearbeitern, Trainingsnutzung und Bearbeitungsort sowie bei kritischen Abläufen zu Ausweich- oder Notbetrieb und Exit.	Änderungsinformationen der Anbieter werden bewertet. Wesentliche Änderungen führen nachweisbar zu einer erneuten Beurteilung und bei Bedarf zu angepassten Konfigurationen. Für kritische Abläufe liegt eine Ausweich- oder Notbetriebslösung vor, deren Tragfähigkeit geprüft ist (Nachweise im BCM).
Risikobasierte technische Kontrollen sind definiert, etwa für Zugriffe, Mandantenkonfiguration und Datenabfluss sowie bei agentischen Systemen für Identitäten, zugelassene Funktionen und Aktionsgrenzen.	Konfigurationen, Berechtigungen und Ausnahmen werden regelmässig überprüft. Ausnahmen sind genehmigt und befristet. Bei agentischen Systemen werden zusätzlich Aktionsprotokolle, Freigabeschwellen und Abschaltmechanismen risikobasiert getestet.
Ein niederschwelliger Meldeweg für Beobachtungen, Bedenken und Vorfälle rund um KI ist eingerichtet und bekannt.	Meldungen gehen tatsächlich ein, werden nachvollziehbar behandelt und führen bei Bedarf zu Anpassungen an Inventar, Regeln oder Konfiguration.
Für wesentliche Anwendungsfälle sind Qualitätskriterien, Testfälle, Grenzen der Verwendung und Anforderungen an menschliche Kontrolle festgelegt.	Tests und geprüfte Ergebnisstichproben belegen die fachliche Eignung. Abweichungen führen zu Korrekturen, Einschränkungen oder zur Ausserbetriebnahme.
Für wesentliche Anwendungsfälle sind Anforderungen an Überwachung, Änderungen und Ausserbetriebnahme definiert.	Überprüfungen finden wie vorgesehen statt. Veränderungen von Modell, Daten, Nutzung oder Risikoprofil führen zu einer erneuten Beurteilung. Nicht mehr benötigte Anwendungsfälle werden kontrolliert ausser Betrieb genommen.

Tabelle 4: Design Effectiveness und Operating Effectiveness in der KI-Governance

Der Wirksamkeitsnachweis ist ohne personenbezogene Verhaltensüberwachung möglich. Als Prüfnachweise eignen sich Stichproben aus Freigabeentscheiden, dokumentierte Anbieteränderungen, Rezertifizierungen von Berechtigungen, Nachweise erfolgter Überprüfungen, geprüfte Ergebnisstichproben, aggregierte und nicht personenbezogene Auswertungen technischer Kanal- und Datenkontrollen und dokumentierte Abweichungen vom Regelfall, bei agentischen Systemen zusätzlich Aktionsprotokolle und Abschalttests. Vorfälle und Beinahe-Vorfälle zeigen ergänzend, wo eine Regel im Alltag nicht greift. Für die individuelle Nutzung bleibt der Nachweis, dass Regeln befolgt werden, allerdings indikativ (vgl. Whitepaper Prüfwerkzeuge dieser Reihe).

Nutzungszahlen des freigegebenen Kanals sind ein nützliches, aber kein beweisendes Signal. Eine auffällig geringe Nutzung trotz erkennbarem fachlichem Bedarf belegt keine Schatten-KI, sollte aber Anlass geben, Bekanntheit, Nutzbarkeit und Freigabegeschwindigkeit des zugelassenen Wegs zu hinterfragen.

8 Leitfragen für Leitungs- und Aufsichtsgremien

Die folgenden Fragen verbinden die beiden Achsen: wie weit die KI-Nutzung entwickelt ist und wie tragfähig deren Steuerung ist (vgl. Abbildung 1). Sie eignen sich für eine Standortbestimmung im Leitungs- oder Aufsichtsgremium und als Einstieg in ein Self-Assessment.

Zur Entwicklung der Nutzung

- Wissen wir, welche KI-Anwendungsfälle in unserer Organisation produktiv genutzt werden, einschliesslich der KI-Funktionen in bereits freigegebenen Anwendungen unserer Dienstleister?
- Welche Integrations- und Autonomiestufen nutzen wir heute faktisch, und erfüllen unsere Anwendungsfälle die Voraussetzungen ihrer Stufe, oder sind sie in diese hineingewachsen?

Zur Tragfähigkeit der Steuerung

- Gibt es einen freigegebenen Weg, der den legitimen Bedarf unserer Mitarbeitenden abdeckt, und wie lange dauert die Freigabe eines neuen Anwendungsfalls?
- Ist die Ergebnisverantwortung je Anwendungsfall benannt und wird sie wahrgenommen, oder liegt sie faktisch bei Funktionen, die den fachlichen Nutzen nicht verantworten?
- Sind Informationspflichten zu Funktionsänderungen, weiteren Bearbeitern, Bearbeitungsort und Trainingsnutzung mit unseren wesentlichen Anbietern vertraglich geregelt, und wissen wir, wie wir kritische Abläufe beim Ausfall eines Anbieters weiterführen?
- Woran würden wir heute erkennen, dass unsere Nutzungsregeln nicht befolgt werden, ohne dabei auf die Überwachung von Mitarbeitenden zurückzugreifen?
- Falls wir agentische Systeme einsetzen oder planen: Verfügen diese über eigene Identitäten, zugelassene Funktionen, Aktionsgrenzen, protokollierte Aktionen und eine geprüfte Abschaltmöglichkeit?
- Welche Nachweise könnten wir heute dafür vorlegen, dass unsere KI-Governance nicht nur dokumentiert ist, sondern im operativen Betrieb tatsächlich funktioniert?

Wer diese Fragen heute nicht beantworten kann, hat damit einen konkreten Ausgangspunkt für das nächste Governance-Review und nicht zwingend ein KI-Problem.

Fazit

KI-Governance braucht keine neue Steuerungslogik. Sie wendet bestehende Steuerungsprinzipien auf ein Objekt an, das sich schneller verbreitet als frühere Technologien und dessen Einführung kaum Spuren in den üblichen Kontrollpunkten hinterlässt, und ergänzt sie dort, wo KI eigene Fragen stellt: Folgenabschätzung, fachliche Qualität, Modellverhalten und eigenständig handelnde Systeme. Schatten-KI zeigt, wo diese Steuerung fehlt.

Wirksam sind weder Verbote allein noch eine personenbezogene Überwachung der Mitarbeitenden. Wirksam sind Transparenz über die tatsächliche Nutzung, klare Regeln entlang der Datenklassifizierung, eine benannte Ergebnisverantwortung, ein tragfähiger freigegebener Weg und die Steuerung jener Risiken, die zusätzlich vertragliche oder technische Kontrollen erfordern. Ob diese Elemente wirken, zeigt erst der Nachweis im Betrieb: an Freigaben, Anbieteränderungen, Ergebnisstichproben und Abweichungen, nicht an Weisungen. Der Nutzen liegt dabei nicht in der Absicherung des Status quo, sondern in der Entwicklungsfähigkeit: Nur eine Organisation, die ihren Umgang mit KI steuern kann, geht den Schritt zu integrierten und agentischen Anwendungen bewusst, statt ihn unbemerkt zu vollziehen.

Ohne Steuerung entscheidet die Organisation nicht, wie weit sie mit KI geht. Sie merkt es erst an den Folgen.

Quellen

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

Bundesrat, «KI-Regulierung: Bundesrat will Konvention des Europarates ratifizieren», Medienmitteilung vom 12. Februar 2025.

Eidgenössische Finanzmarktaufsicht FINMA, Aufsichtsmitteilung 08/2024 zu Governance und Risikomanagement beim Einsatz künstlicher Intelligenz, 18. Dezember 2024.

Europarat, Rahmenübereinkommen über künstliche Intelligenz und Menschenrechte, Demokratie und Rechtsstaatlichkeit (CETS Nr. 225), 2024.

ISO/IEC 42001:2023, Information technology — Artificial intelligence — Management system; ergänzend ISO/IEC 23894:2023, ISO/IEC 42005:2025, ISO/IEC 27001:2022 und ISO 22301:2019.

Obligationenrecht (OR, SR 220), Art. 328b.

OWASP GenAI Security Project, OWASP Top 10 for LLM Applications, Fassung 2025.

Schweizerisches Strafgesetzbuch (StGB, SR 311.0), Art. 162, 320 und 321.

Verordnung 3 zum Arbeitsgesetz (ArGV 3, SR 822.113), Art. 26.

Verordnung über den Datenschutz (Datenschutzverordnung, DSV, SR 235.11).

Verordnung (EU) 2024/1689 (KI-Verordnung), ABl. L, 2024/1689, 12.7.2024, geändert durch die Verordnung (EU) 2026/1744 (Digital-Omnibus-Verordnung zur KI), ABl. L, 2026/1744, 24.7.2026.

Hinweis

Dieses Whitepaper gibt ausschliesslich meine persönliche fachliche Einschätzung wieder und steht nicht für eine Behörde oder Aufsicht. Es darf unverändert weitergegeben und mit Quellenangabe zitiert werden. Kommerzielle Verwertung und Bearbeitungen bedürfen meiner vorherigen Zustimmung.



Fragen und Austausch willkommen

carsten@carsten-beck.ch

Alle Beiträge der Reihe

carsten-beck.ch/publikationen

linkedin.com/in/carstenbeck-grc