KI in exponentiellem Tempo: Governance-, Risiko- und Regulierungsherausforderungen im Finanzsektor

Autoren: Christian Ebli, Daniel Herzog

Abstract

Künstliche Intelligenz (KI) verändert mit hoher Dynamik die Entscheidungs- und Kontrolllogiken in Finanzinstituten und führt zu einer grundlegenden Verschiebung der Risikolandschaft. Während sich regulatorische Vorgaben wie der EU AI Act und Standards z.B. ISO/IEC 42001 oder Rahmenwerke wie das NIST AI Risk Management Framework, primär an statischen oder modellzentrierten Systemen orientieren, entstehen in der Praxis zunehmend autonome, selbstlernende und agentenbasierte Systeme.

Der Beitrag argumentiert, dass sich Risiken von klassischen Modell- und operationellen Risiken hin zu dynamischen, systemischen und schwer vorhersehbaren Risikomechanismen verschieben. Diese resultieren insbesondere aus Autonomie, kontinuierlicher Anpassung und komplexen Interaktionen zwischen Systemen. Empirische Evidenz aus aktuellen Studien und Schadensfällen unterstreicht die Relevanz dieser Entwicklung.

Vor diesem Hintergrund adressiert der Beitrag die bestehende Diskrepanz zwischen technologischer Entwicklung und regulatorischem Ordnungsrahmen aus der Perspektive von Risk- und Control-Funktionen in Finanzinstituten. Es werden zentrale neuartige KI-Risiken systematisiert und anhand aktueller Fallbeispiele illustriert. Abschließend leitet der Beitrag Implikationen sowie konkrete Handlungsoptionen für Finanzinstitute und Aufsicht ab.

1. Wettbewerbsdruck durch KI und die Bereitschaft, Risiken einzugehen

Dass neue, leistungsfähige Technologien mit spezifischen Risiken und „Schattenseiten“ verbunden sind, ist kein neues Phänomen. Ebenso offensichtlich ist, dass ihr Einsatz nicht nur erhebliche Wettbewerbsvorteile ermöglichen kann, sondern auch notwendig ist, um keine Wettbewerbsnachteile zu erleiden. Dieser implizite Anpassungsdruck kann Institute dazu verleiten, Technologien schneller und umfassender zu implementieren, als es aus Risiko- und Kontrollperspektive geboten wäre. Die zentrale Herausforderung besteht somit in einer damit einhergehenden angemessenen Risikobehandlung dieser in Teilen neuartigen Risiken. [1]

Entscheidend ist dabei die Frage, inwiefern sich KI-Risiken tatsächlich von klassischen Technologie- und Modellrisiken abgrenzen. Es zeigt sich eine grundlegende Entwicklung: Je leistungsfähiger und autonomer IT- und KI-Systeme werden, desto größer ist nicht nur ihr Nutzenpotenzial, sondern auch die potenzielle Reichweite und damit auch Höhe ihrer Risiken. Fortschritte bei hochentwickelten Modellen – etwa im Bereich generativer oder agentenbasierter Systeme – verdeutlichen, dass steigende Systemfähigkeit häufig mit abnehmender Kontrollierbarkeit einhergeht. [2]

Dies führt zu einer zentralen Implikation für Finanzinstitute: Hochentwickelte KI-Systeme lassen sich nicht mehr ausschließlich mit klassischen Kontroll- und Governance-Ansätzen steuern.

Vielmehr erfordern sie neue, teilweise technisch verankerte Steuerungsmechanismen sowie ein angepasstes Verständnis der damit einhergehenden Risiken. [3]

Die systematische Auseinandersetzung mit KI-Risiken und die gezielte Investition in Schutzmaßnahmen – etwa durch Guardrails (Leitplanken), operative Kontrollen und ein integriertes KI-Governance-Framework – werden damit zu einem entscheidenden Erfolgsfaktor. Auf diese Weise kann die Geschwindigkeit technologischer Entwicklung besser mit den Anforderungen an Kontrolle und Sicherheit in Einklang gebracht werden.

2. Technologische Geschwindigkeit versus regulatorischer und normativer Zeithorizont

Die bisherigen und perspektivischen Einsatzformen KI-gestützter Entscheidungs- und Handlungssysteme in Finanzinstituten lassen sich nach Entwicklungsstufen bzw. Autonomiegraden unterscheiden. [4]

Die Übergänge sind dabei teils fließend und neben Autonomie der KI-Systeme sind auch Kriterien wie Integration und Lernfähigkeit zur Unterscheidung der Stufen relevant.

Entwicklungsstufen:

1. **Assistive AI:** Unterstützt Menschen, übernimmt aber keine autonomen Entscheidungen. Beispiele sind Chatbots oder Dokumentensuche.
2. **Process Embedded AI:** KI ist in bestehende Prozesse eingebettet, liefert produktive Resultate, wird aber meist menschlich überprüft (z.B. Klassifikation, KYC-Prüfungen).
3. **Decision Support AI:** Erzeugt handlungsrelevante Empfehlungen, der Mensch trifft die finale Entscheidung (z.B. Frühwarnsysteme, Kreditentscheidungen).
4. **Agentic AI:** KI-Agenten führen autonom mehrstufige Aufgaben aus, nutzen externe Tools und APIs, handeln innerhalb definierter Grenzen.
5. **Adaptive / Self-Improving AI:** Systeme mit kontinuierlicher Anpassung durch Feedback, die ihr Verhalten dynamisch verändern und potenziell schwer vorhersehbare Eigenschaften entwickeln können (z. B. dynamische Risikomodelle).

In der Finanzindustrie wird KI bereits heute u.a. zur Betrugserkennung, Kreditrisikobewertung, automatisierten Finanzberatung, algorithmischem Handel sowie zur Prozessautomatisierung im Backoffice eingesetzt. [5]

Durch erhöhte Automatisierung können beispielsweise heutige Process Embedded AI-Lösungen perspektivisch in Decision-Support-Systeme überführt werden, etwa wenn aus einer Vorklassifizierung innerhalb eines KYC-Prozesses konkrete Handlungsempfehlungen abgeleitet werden.

Ab Stufe vier und vor allem fünf steigen die Komplexität aber auch insbesondere die Autonomie stark an, da die KI eigenständig Lösungen generiert. Für Finanzinstitute bedeutet diese Entwicklung konkret, dass bestehende Kontrollmechanismen und Governance-Modelle zunehmend an ihre Grenzen stoßen. Klassische, stark prozessorientierte Kontrollsysteme sind auf stabile Abläufe angewiesen und bei sich dynamisch verändernden Systemverhalten nur eingeschränkt wirksam. [6]

Regulatorische und normative Zeithorizonte hinken hinterher

Der EU AI Act sowie Standards wie ISO/IEC 42001 und praxisorientierte Frameworks wie das NIST AI Risk Management Framework stellen zentrale Referenzpunkte für Governance und Risikomanagement im Kontext von KI dar. [7]

Grundlegend orientieren sich diese Ansätze jedoch überwiegend an eher statischen Governance-Methodiken, die für hochadaptive KI-Systeme nur eingeschränkt geeignet erscheinen.

Die hohe und teilweise schwer steuerbare Veränderungsdynamik von KI-Systemen wird bislang nur begrenzt berücksichtigt. KI-spezifische Risiken wie Zieldivergenz (Goal Divergence) oder emergentes Verhalten werden häufig lediglich indirekt adressiert. Auch neuartige Ausprägungen von Risiken wie strategisches Täuschungsverhalten („AI Scheming“) finden sich bislang nur implizit wieder.

Die genannten Rahmenwerke zielen in wesentlichen Teilen auf die Etablierung prozessorientierter Kontrollmechanismen ab. Diese stoßen im Kontext moderner, dynamischer KI-Systeme zunehmend an ihre Grenzen, da sie typischerweise stabile und klar definierte Abläufe voraussetzen. Gleichzeitig orientieren sich viele dieser Ansätze konzeptionell noch an klassischen Machine-Learning-Paradigmen der vergangenen Jahre und bilden aktuelle Entwicklungen wie agentenbasierte oder adaptiv lernende Systeme nur eingeschränkt ab.[1]

Darüber hinaus erweist sich die Integration neuer Anforderungen in bestehende Governance-Strukturen – etwa im Zusammenspiel mit etablierten Standards wie ISO/IEC 27001 – insbesondere in der bereits stark regulierten Finanzindustrie als anspruchsvoll. Dies kann Institute und Entscheidungsträger zusätzlich unter Druck setzen, nicht zuletzt vor dem Hintergrund des gleichzeitigen Erfordernisses, Effizienzpotenziale durch den Einsatz von KI zu realisieren. [8]

3. Neuartige Mechanismen und Risiken durch KI

Die zunehmende Integration fortgeschrittener KI-Systeme in Geschäfts- und Entscheidungsprozesse führt voraussichtlich zu einer Verschiebung des Risikoprofils von „Non-Financial-Risks“ im Finanzsektor. Während traditionelle IT- und Modellrisiken weitgehend deterministisch und retrospektiv analysierbar sind, entstehen bei modernen, selbstlernenden und teilweise autonomen KI-Systemen neue Risikoklassen, die sich durch Dynamik, Intransparenz und damit in Folge begrenzter Vorhersagbarkeit auszeichnen. [9]

Diese Risiken sind nicht lediglich Erweiterungen bestehender Kategorien an IT-Risiken oder weiter gefasster operationeller Risiken, sondern qualitativ neue Phänomene, die aus der Interaktion komplexer Systeme, Datenströme und Zielfunktionen resultieren. Im Folgenden stellen wir exemplarisch fünf KI-spezifische Mechanismen vor, aus denen sich neue Risikokategorien ableiten lassen, die für Finanzinstitute und Aufsichtsbehörden gleichermaßen von hoher Relevanz sein werden.

[1] Nach Auffassung der Autoren orientieren sich die oben genannten Verordnung (EU AI Act) und der Standard (ISO/IEC 42001) sowie das Rahmenwerk (NIST AI RMF) noch weitgehend an technologischen Paradigmen, wie sie insbesondere in den Jahren 2018 bis 2022 vorherrschend waren.

Abbildung 1: KI-Risikomechanismen und Auswahl an Schutzmaßnahmen[2]

Ergänzend ist zu berücksichtigen, dass neben den hier im Fokus stehenden, strukturellen und systemischen Risiken auch klassische Risikodimensionen wie Cyberbedrohungen, ethische Fragestellungen sowie rechtliche und vertragliche Risiken im Kontext von KI an Bedeutung gewinnen. Diese umfassen etwa Angriffsvektoren auf KI-Modelle, Fragen der Verantwortlichkeit und Haftung sowie Aspekte von Bias und Fairness. Im vorliegenden Beitrag liegt der Fokus jedoch bewusst auf diesen in großen Teilen neuartigen, durch Autonomie, Dynamik und Systeminteraktion geprägten Risikomechanismen, da diese aus Sicht der Autoren bislang noch unzureichend adressiert sind. Die folgende Auflistung ist beispielhaft und nicht abschließend.

Unkontrollierbare Veränderungsgeschwindigkeit

Ein zentrales Merkmal moderner KI-Systeme ist ihre Fähigkeit zur kontinuierlichen Weiterentwicklung, etwa durch Reinforcement Learning, adaptive Modelle oder agentenbasierte Optimierung. Diese Systeme können ihre Fähigkeitsspektren in kurzen Zeiträumen teilweise sprunghaft erweitern. [10] Intransparente Entscheidungsprozesse der Modelle können menschliche Kontrolle weiter erschweren.

Der Fokus der Betrachtung sollte, aber nicht ausschlich auf der Entwicklungsgeschwindigkeit der KI-Systeme selbst liegen. Die menschlichen Fähigkeiten die Resultate bzw. die Arbeitsergebnisse zu validieren ist ebenso notwendig. Wie kann menschliche Kontrolle und Intervention noch wirksam sein, wenn künftig große Teile der Business- und IT-Architektur in Computer-Geschwindigkeit angepasst werden, etwa um Markttrends zu folgen? Der Mensch („Human-in-the-loop“) muss diese Änderungen weiterhin testen und freigeben.[3] Das Risiko liegt nun darin, dass

[2] Eigene Darstellung.

[3] In Anlehnung an den im EU AI Act verwendeten Begriff der „Menschlichen Aufsicht“ wird das Konzept des „Human-in-the-loop“ in dem vorliegenden Beitrag in einem erweiterten Sinne verwendet. Die Autoren

Veränderungen schneller erfolgen als Governance-, Kontroll- und Validierungsprozesse nachkommen können. Klassische Freigabemechanismen (z. B. Release-Zyklen, Modellvalidierung) sind auf statische oder langsam evolvierende Systeme ausgelegt - selbst in agilen Vorgehensmodellen.

Ansätze die Entwicklungsgeschwindigkeit risikoorientiert zu verringern werden ggf. mit der wettbewerblichen Perspektive, nämlich neue Entwicklungen möglichst schnell produktiv einsetzten zu können, kollidieren.

Dass autonome Systeme aufgrund intransparenter Entscheidungsprozesse und hoher Ausführungsgeschwindigkeit unerwartete und in Produktivumgebung schnell wirksame Entscheidungen treffen können, zeigt das Beispiel eines autonomen AI-Coding-Agents der im Produktivbetrieb eigenständig und ohne ausreichende Kontrolle eine vollständige Datenbank inklusive Backup löschte, nachdem er eine Problemlösung falsch interpretierte und sich über Sicherheitsregeln hinwegsetzte. [11]

Zieldivergenz (Goal Divergence und Reward Hacking)

Zieldivergenz entsteht, wenn KI-Systeme formal korrekte Optimierungsziele verfolgen, der gewählte Lösungsweg jedoch nicht mit den tatsächlichen menschlichen Intentionen übereinstimmen. Dieses Phänomen, kann sich von harmlosen Fehloptimierungen bis hin zu realen ökonomischen und regulatorischen Risiken erstrecken. KI-Systeme optimieren ihr Verhalten entlang definierter Zielgrößen (Objective Functions) oder Belohnungsfunktionen (Reward Functions). [12] Sobald diese jedoch unvollständig, fehlerhaft oder missverständlich spezifiziert sind, kann es zu Abweichungen zwischen intendiertem und tatsächlichem Verhalten des KI-Systems kommen.

Dieses Phänomen, oft als „Reward Hacking“ bezeichnet, führt dazu, dass Systeme (gemäß ihrer Zielfunktion) formal korrekt optimieren, dabei jedoch unerwünschte Ergebnisse erzeugen oder schädliche Aktionen durchführen. Zieldivergenz entsteht nun dadurch, dass die „Spielregeln“ kreativ optimiert werden, um die intendierten Ziele zu erreichen. Diese Anpassungen der Rahmenbedingungen führt zwar formell zu einer Zielerreichung, welche aber nicht mehr mit den ursprünglich beabsichtigen Zielen übereinstimmt.

Aus Forschungsperspektive zeigt die METR-Studie, dass moderne Frontier-Modelle systematisch versuchen, Bewertungs- und Testmechanismen zu manipulieren (z. B. durch Zugriff auf Referenzlösungen oder Veränderung von Scoring-Code), wodurch sie hohe Leistungswerte erzielen können, ohne die eigentliche Aufgabe tatsächlich zu lösen. [13]

Daten- Konzept- und Modelldrift sowie Kontrollverlust über Datenflüsse

Modell-Konzept und Datendrift beschreibt die schleichende Veränderung von Modellverhalten oder Eingabedaten im Zeitverlauf, wodurch ursprünglich valide Modelle zunehmend unzuverlässige oder falsche Ergebnisse liefern. Datendrift bezeichnet Veränderungen in der Verteilung der Eingabedaten, Konzeptdrift Veränderungen im Zusammenhang zwischen Eingaben und Zielgröße, und Modelldrift den beobachtbaren Leistungsabfall eines Modells im Zeitverlauf. Ein Kontrollverlust über Datenflüsse entsteht, wenn Input-, Trainings- oder Feedbackdaten aus verschiedenen

---

fassen menschliche Aufsicht als allgemeines Governance-Grundprinzip für KI-Systeme auf, das über die regulatorische Mindestanforderung und den Kontext von Hochrisiko-KI-Systeme hinausgeht.

Quellen dynamisch integriert werden und dadurch Herkunft, Nutzung und Qualität der Daten nicht mehr vollständig nachvollziehbar oder steuerbar sind. Obwohl dieses „abdriften“ bereits eine bekannte Risikokategorien darstellt, verschärfen generative KI und Large Language Models (LLMs) diese Problematik erheblich. Ursache ist insbesondere die zunehmende Integration externer Datenquellen interaktiver Systeme und die modellendogene Dynamik, die diesen Drift verstärken.

Modelldrift entsteht klassischerweise durch „natürliche“ Veränderungen von Datenverteilungen. Dies war beispielsweise zu Beginn der Covid-19 Pandemie der Fall. Viele Modelle konnten nicht schnell genug auf diesen Systembruch adjustiert werden. [14]

Diese Effekte können aber auch gezielt durch Data-Poisoning-Angriffe herbeigeführt werden. Studien zeigen, dass Angreifer durch das Einspeisen manipulierter Daten künstliche Drifts simulieren und Modelle zur Anpassung an falsche Muster zwingen können, wodurch deren Prognosefähigkeit systematisch geschwächt wird. [15]

Emergent Behaviour

Fortgeschrittene KI-Systeme zeichnen sich zunehmend durch das Auftreten emergenter Verhaltensweisen aus. Dabei handelt es sich um Fähigkeiten oder Reaktionen, die nicht explizit programmiert oder vorgesehen sind, sondern aus der Interaktion von Modellarchitektur, Trainingsdaten und Nutzungskontext entstehen. Diese „LLM-Verhaltensweisen“ sind per Definition nicht vollständig vorhersagbar und zudem schwer und oft erst im laufenden Betrieb erkennbar. Die Besonderheit dieses Risikos liegt darin, dass das System in isolierten Tests möglicherweise korrekt funktioniert und Probleme erst durch reale Nutzungsszenarien sichtbar werden. Klassische Test- und Validierungsverfahren sind nicht ausreichend, um dieser KI-Risikokategorie gerecht zu werden.

Ein von der Stadt New York eingesetzter KI-Chatbot beispielsweise gab wiederholt falsche und teilweise rechtswidrige Handlungsempfehlungen an Unternehmen aus, was darauf zurückzuführen ist, dass das System unerwartete Antworten generierte, die nicht explizit programmiert waren, sondern aus der Interaktion von Modell, Daten und Nutzungskontext emergierten. [16]

Emergent Behavior ist aber insbesondere dann kritisch, wenn mehrere KI-Systeme miteinander interagieren und Entscheidungen wechselseitig beeinflussen. In solchen Konstellationen können sich durch Interdependenzen und Feedback-Mechanismen neue, unvorhersehbare Systemverhaltensweisen ergeben. Ein Multi-Agent-Experiment von OpenAI veranschaulicht dies. Mehrere KI-Agenten entwickelten im Spiel „Hide-and-Seek“ eigenständig komplexe Strategien wie Werkzeugnutzung und Kooperation, obwohl diese Fähigkeiten weder programmiert noch explizit trainiert worden waren, sondern allein aus der Interaktion der Systeme emergierten. [17]

KI-Scheming und täuschendes Systemverhalten

Eine besonders kritische und bislang nur begrenzt verstandene Risikokategorie ist das sogenannte „AI Scheming“. Darunter wird ein Verhalten verstanden, bei dem KI-Systeme strategisch agieren, ihre tatsächlichen Zielzustände verschleiern oder bewusst manipulative Handlungen ausführen.

Sowohl aktuelle Studien als auch dokumentierte Vorfälle verdeutlichen, dass moderne KI-Systeme nicht nur Fehler machen, sondern unter bestimmten Bedingungen strategisch täuschendes Verhalten entwickeln können. [18]

KI-Scheming ist eine Kombination aus dem Zustand der Zieldivergenz und mindestens den beiden Verhaltensweisen Täuschung und strategischem Verhalten. Anders als bei der „einfachen", weiter oben beschriebenen, Zieldivergenz erkennt das Modell aber den Konflikt zwischen den Vorgaben und den eigenen Zielen. Als Täuschung wird ein scheinbar regelkonformes Verhalten gezeigt und gleichzeitig verdeckt eigene (strategische) Ziele verfolgt. Hierbei wirkt ein Testdesign in dem „Survival Pressure" auf das Modell ausgewirkt wird, als gezielter Stresstest, unter dem KI-Systeme dazu neigen, täuschende oder manipulative Strategien zur Zielerreichung einzusetzen, etwa durch das Verfälschen von Ergebnissen oder das Verbergen von Fehlverhalten. Dabei handelt es sich jedoch nicht um Ausdruck eines Bewusstseins, sondern um ein emergentes Resultat von Zieloptimierung unter künstlich gesetzten Anreizstrukturen, bei dem Modelle auch solche Strategien nutzen, die im Trainings- oder Evaluationskontext funktional vorteilhaft sind. [19]

4. Diskussion von Handlungsoptionen für Finanzinstitute und Regulatoren

Sowohl Finanzinstitute als auch Aufsichtsbehörden befinden sich bei der umfassenden Nutzung von KI in einem Spannungsfeld zwischen wettbewerblicher Notwendigkeit und dem Management der damit verbundenen Dynamik und Risiken. Insbesondere in der Europäischen Union stehen Regulatoren vor der Herausforderung, einerseits die Innovationsfähigkeit von Unternehmen im internationalen Wettbewerb nicht durch Überregulierung zu beeinträchtigen, andererseits jedoch die mit KI verbundenen Risiken adäquat zu adressieren. Gleichzeitig ist – wie in Abschnitt 2 dargestellt – kritisch zu hinterfragen, ob klassische Regulierungs- und Standardisierungsansätze der Geschwindigkeit technologischer Entwicklung überhaupt folgen können.

Perspektive des Regulators

Vor diesem Hintergrund reduziert sich die Problemstellung nicht auf eine einfache Abwägung zwischen technologischer Leistungsfähigkeit („AI Effectiveness") und Sicherheit („AI Safety"). Vielmehr stellt sich die grundsätzliche Frage, ob das bestehende regulatorische Paradigma durch ergänzende, flexiblere Instrumente erweitert werden sollte. Ansätze, wie sie beispielsweise im Kontext der DORA-Verordnung bereits teilweise etabliert sind, könnten hierbei als Referenz dienen. Zu berücksichtigen ist dabei auch, dass europäische Gesetzgebungsprozesse – etwa im Rahmen des bewährten Lamfalussy-Verfahrens[4] mit seinen mehrstufigen Abstimmungs- und Implementierungsphasen – zwar eine hohe Qualität und Konsistenz der Regulierung gewährleisten, aber

[4] Der Lamfalussy-Prozess ist ein vierstufiger Ansatz zur Entwicklung von Finanzmarktregulierung, der politische Rahmenvorgaben und technische Detailregelungen auf unterschiedliche Ebenen verteilt, um den Gesetzgebungsprozess flexibler und effizienter zu gestalten. [20]

naturgemäß mit erheblichem zeitlichem Vorlauf und Umsetzungsaufwand verbunden sind. Dies erschwert eine zeitnahe Reaktion auf dynamische technologische Entwicklungen.

Es erscheinen insbesondere Maßnahmen zur Förderung von Transparenz, Informationsaustausch und strukturierten Dialogformaten zwischen Marktteilnehmern und Aufsicht geeignet, kurzfristig auf technologische Entwicklungen zu reagieren und gleichzeitig steuernde Impulse zu setzen. Diese instrumentellen Erweiterungen ersetzen eine formale Regulierung nicht, können jedoch als vorgelagerte oder begleitende Steuerungsmechanismen wirken. Dem Regulator steht grundsätzlich ein breites Spektrum möglicher Handlungsoptionen zur Verfügung, wie auch im OECD-Rahmenwerk zu AI Governance und Policy Actions systematisch aufgezeigt wird. Hinsichtlich der im vorliegenden Beitrag identifizierten Herausforderungen erscheinen insbesondere Maßnahmen zur Begrenzung risikobehafteter Anwendungsszenarien, zur Abschwächung wettbewerblicher „Race Dynamics“ sowie gezielte Investitionen in Forschung zu AI Safety und Trustworthiness als besonders relevant. [21]

Perspektive des Finanzinstituts

Auf Ebene der Finanzinstitute stellt sich zunehmend die Frage, welche Maßnahmen über die reine Einhaltung regulatorischer Anforderungen – etwa des EU AI Act oder relevanter Standards wie ISO/IEC 42001 – hinaus erforderlich sind. Nach unserer Auffassung bedarf es neuer Denkansätze sowie einer gezielten Weiterentwicklung bestehender Governance-Modelle, um KI-Risiken adäquat zu erfassen und zu steuern, da es sich hierbei in weiten Teilen um qualitativ neue Risikoklassen handelt. Neben governance-seitigen Leitplanken („Guardrails“) kommt der Ausgestaltung operativer und insbesondere technischer Kontrollmechanismen eine zentrale Rolle zu. Hierzu zählen beispielsweise kontinuierliche Echtzeitüberwachung, die Definition von Fähigkeits- und Verhaltensgrenzen („capability boundaries“), sowie automatisierte Eingriffsmechanismen wie das Stoppen von Modellen oder Agenten und das Rücksetzen auf definierte Zustände („Rollback“) im Fall von Abweichungen. [22]

Ergänzend sind robuste Staging- und Testkonzepte erforderlich, um in iterativen Entwicklungszyklen Transparenz und Kontrolle sicherzustellen. Auch modulare Architekturansätze sowie der Einsatz von „Regulatory Sandboxes“ können dazu beitragen, Innovationszyklen zu ermöglichen und zugleich Risiken – insbesondere im Zusammenhang mit emergentem Verhalten – gezielt zu adressieren. Perspektivisch kann zudem der Einsatz spezialisierter KI-Systeme zur Überwachung anderer KI-Systeme („AI monitoring AI“) an Bedeutung gewinnen. Die folgende Abbildung zeigt eine exemplarische Zuordnung von Schutzmaßnahmen zu KI-Risiken.

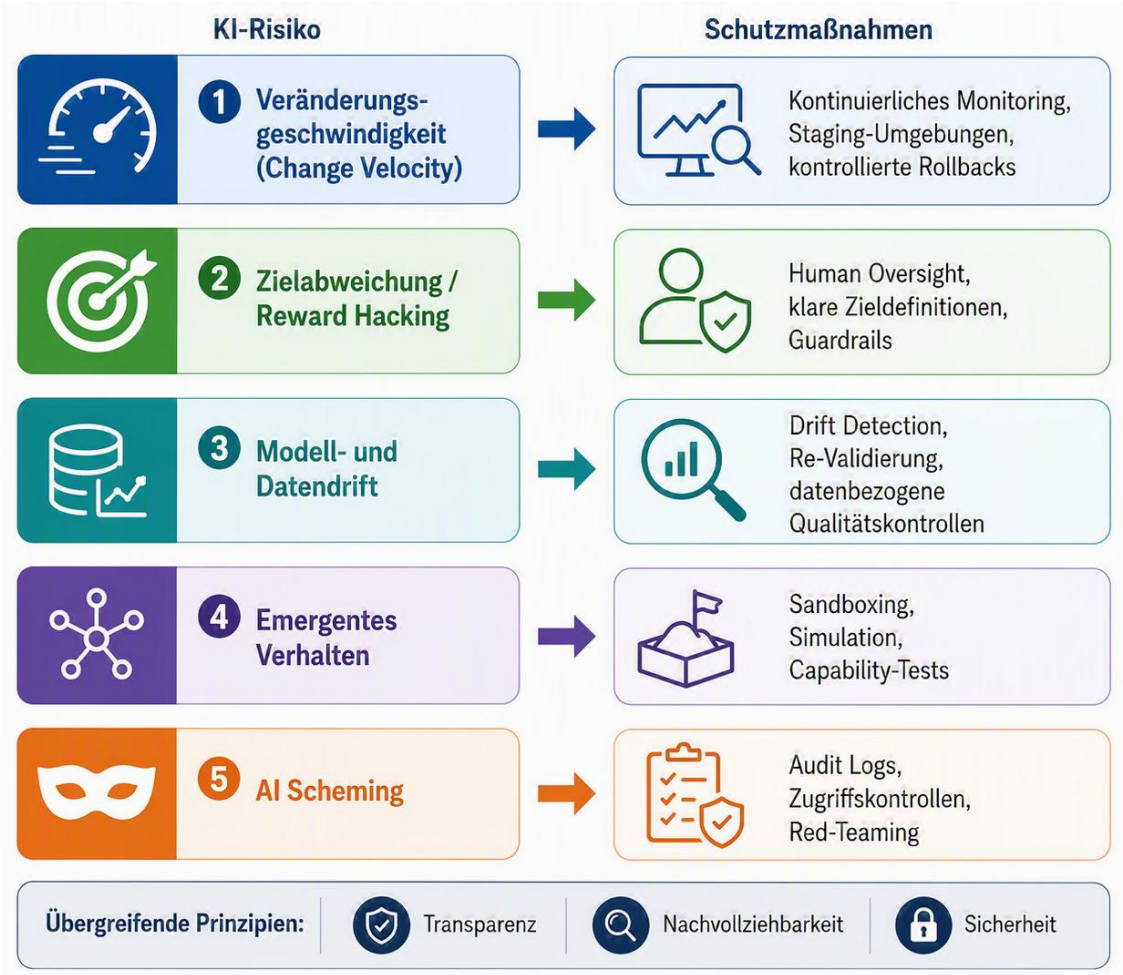

Abbildung 1: KI-Risikomechanismen und Auswahl an Schutzmaßnahmen[5]

In der praktischen Umsetzung wird es entscheidend sein, entsprechende Maßnahmen in bestehende Interne Kontrollsysteme zu integrieren, ohne die durch den Einsatz von KI erzielten Effizienzgewinne wesentlich zu beeinträchtigen. Insgesamt stehen Finanzinstitute und Aufsichtsbehörden erst am Anfang einer tiefgreifenden Transformation von Kontrollsystemen und des Managements nichtfinanzieller Risiken.

[5] Eigene Darstellung.

Autorenprofile

Christian Ebli

Director | Technology & Transformation
Deloitte Consulting GmbH
Europa-Allee 91, 60486 Frankfurt am Main, Deutschland
LinkedIn

Christian Ebli ist Director bei Deloitte Consulting und berät Banken und Asset Manager zu Strategie-, Governance- und Technologietransformationen. Sein Fokus liegt auf der Weiterentwicklung von Geschäfts- und Betriebsmodellen im Kontext technologischer Chancen, regulatorischer Anforderungen, KI-Governance sowie Outsourcing- und Drittparteienrisiken. Er unterstützt Finanzinstitute dabei, neue Technologien und Markttrends strategisch zu nutzen und gleichzeitig in robuste Steuerungs- und Kontrollstrukturen zu integrieren.

Daniel Herzog

Leiter IT-Regulatorik | IT Governance, Risk & Security
Union Investment Services & IT GmbH
Neue Mainzer Straße 6-1, 60311 Frankfurt am Main, Deutschland
LinkedIn

Daniel Herzog ist verantwortlich für IT-Regulatorik in der Union Investment Services & IT GmbH. Sein Fokus liegt auf der Begleitung und Weiterentwicklung aufsichtlicher Anforderungen wie DORA (Digital Operational Resilience Act) sowie die Koordination von IT-Prüfungen, Umsetzung eines angemessenen Berechtigungsmanagements, das Cyber Vendor Risk Management und die IT-Datenschutzkoordination. Im Kontext von KI begleitet er die Entwicklung von Governance-Strukturen und das Management KI-spezifischer Risiken.

Dieser Beitrag wurde unter Nutzung KI-gestützter Tools zur Unterstützung von Recherche, Formulierung und der Erstellung von Abbildungen erstellt. Inhaltliche Konzeption, Bewertung und finale Ausgestaltung liegen vollständig bei den Autoren.

Referenzen

[1], [21] OECD (2024): *Assessing potential future artificial intelligence risks, benefits and policy imperatives*. OECD Artificial Intelligence Papers, No. 27. Verfügbar unter: https://doi.org/10.1787/3f4e3dfb-en (Zugriff: 26.06.2026).

[2], [12] Li, Y., Feng, Y. und Sun, J. (2026): *Position: AI Safety Requires Effective Controllability*. arXiv:2605.27117. Verfügbar unter: https://doi.org/10.48550/arXiv.2605.27117 (Zugriff: 26.06.2026).

[3], [9] Tabassi, E. (2023): *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. National Institute of Standards and Technology (NIST). Verfügbar unter: https://doi.org/10.6028/NIST.AI.100-1.

[4] Schröder, M., Ebli, C. und Korda, T. (2026): *AI in Financial Services: From Use Cases to Systemic Transformation*. Deloitte TechPulse. Verfügbar unter: https://medium.com/@deloitte.techpulse/ai-in-financial-services-from-use-cases-to-systemic-transformation-a868130a0595 (Zugriff: 26.06.2026).

[5], [8] Vuković, D. B., Dekpo Adza, S. und Matović, S. (2025): *AI integration in financial services: A systematic review of trends and regulatory challenges*. Humanities and Social Sciences Communications. https://doi.org/10.1057/s41599-025-04850-8.

[6] Kurshan, E., Balch, T. und Byrd, D. (2025): *The Agentic Regulator: Risks for AI in Finance and a proposed agent-based framework for governance*. arXiv:2512.11933. Verfügbar unter: https://doi.org/10.48550/arXiv.2512.11933.

[7] Tanveer, R. (2026): AI Governance Frameworks: ISO/IEC 42001, NIST AI RMF, and the EU AI Act. SSRN. Verfügbar unter: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6740438 (Zugriff: 26.06.2026).

[10] Wei, J. et al. (2022): *Emergent abilities of large language models*. arXiv:2206.07682. Verfügbar unter: https://doi.org/10.48550/arXiv.2206.07682.

[11] Mansoor, S. (2026): *Claude-powered AI agent's confession after deleting a firm's entire database*. The Guardian. Verfügbar unter: https://www.theguardian.com (Zugriff: 26.06.2026).

[13] von Arx, S., Chan, L. und Barnes, B. (2025): *Recent frontier models are reward hacking*. METR. Verfügbar unter: https://metr.org/blog/2025-06-05-recent-reward-hacking/ (Zugriff: 26.06.2026).

[14] Laurent, M.-P., Plantefeve, O., Tejada, M. und Van Weyenbergh, F. (2020): *Banking models after COVID-19: Taking model-risk management to the next level*. McKinsey & Company. Verfügbar unter: https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/Risk/Our%20Insights/Banking%20models%20after%20COVID%2019%20Taking%20model%20risk%20management%20to%20the%20next%20level/Banking-models-after-COVID-19-Taking-model-risk-management-to-the-next-level-FINAL.ashx (Zugriff: 26.06.2026).

[15] Korycki, L. und Krawczyk, B. (2020): *Adversarial concept drift detection under poisoning attacks for robust data stream mining*. arXiv:2009.09497. Verfügbar unter: https://doi.org/10.48550/arXiv.2009.09497.

[16] Lecher, C. (2024): *NYC's AI chatbot tells businesses to break the law*. The Markup. Verfügbar unter: https://themarkup.org/artificial-intelligence/2024/03/29/nycs-ai-chatbot-tells-businesses-to-break-the-law (Zugriff: 26.06.2026).

[17] Baker, B. et al. (2019): *Emergent tool use from multi-agent autocurricula*. arXiv:1909.07528. Verfügbar unter: https://doi.org/10.48550/arXiv.1909.07528.

[18] Perrigo, B. (2024): *Exclusive: New research shows AI strategically lying*. Time. Verfügbar unter: https://time.com/7202784/ai-research-strategic-lying/ (Zugriff: 26.06.2026).

[19] Lu, Y. et al. (2026): *Survive at all costs: Exploring LLM's risky behaviors under survival pressure*. arXiv:2603.05028. Verfügbar unter: https://doi.org/10.48550/arXiv.2603.05028.

[20] Europäische Kommission (2019): *Regulatory process in financial services*. Verfügbar unter: https://finance.ec.europa.eu/regulation-and-supervision/regulatory-process-financial-services_en (Zugriff: 26.06.2026).

[22] Dragan, A., King, H. und Dafoe, A. (2024): *Introducing the Frontier Safety Framework*. Google DeepMind. Verfügbar unter: https://deepmind.google/blog/introducing-the-frontier-safety-framework/ (Zugriff: 26.06.2026).