BaFin

Bundesanstalt für
Finanzdienstleistungsaufsicht

Orientierungshilfe zu IKT-Risiken beim Einsatz von KI in Finanzunternehmen

Inhaltsverzeichnis

I. Einleitung 4

1. Begriffsbestimmung eines KI-Systems 6
2. Gegenstand der Orientierungshilfe 6
3. Aufbau der Orientierungshilfe 8

II. IKT-Risikomanagement von KI 9

1. IKT-Risiken aus der Nutzung von KI 9
2. Governance und Organisation 9
3. IKT-Risikomanagementrahmen nach DORA 11

III. Bereitstellung von KI: Entwickeln und Testen 13

1. Softwareentwicklung 13
2. Testen von KI 15

IV. Betrieb und Stilllegung von KI 17

1. Prozesse für den Betrieb und die Deinstallation von KI-Systemen 17
2. Cloud-Spezifika beim Betrieb von KI 19

V. Cyber- und Datensicherheit 22

1. Cybersicherheit 22
2. Datensicherheit 24
3. Meldung schwerwiegender IKT-bezogener Vorfälle 26

VI. Schlussbetrachtung und Ausblick 27

Fallstudie—Betrieb eines LLM-basierten KI-Assistenten 28

I. Einleitung

Diese Orientierungshilfe ist eine nicht verpflichtende Hilfestellung. Sie soll Finanzunternehmen (insbesondere CRR-Institute und Solvency II-Versicherungsunternehmen) dabei unterstützen, beim Einsatz von Künstlicher Intelligenz (KI) einschlägige regulatorische Anforderungen aus dem Digital Operational Resilience Act (DORA)[1] umzusetzen.[2]

Betrachtet wird insbesondere das IKT-Risikomanagement und das IKT-Drittparteienrisikomanagement, inklusive der Delegierten Verordnung zum IKT-Risikomanagement[3] (RTS RMF) sowie der Delegierten Verordnung zur Untervergabe von IKT-Dienstleistungen zur Unterstützung kritischer oder wichtiger Funktionen (RTS Untervergabe)[4].

Die Orientierungshilfe richtet sich somit insbesondere an diejenigen von der BaFin beaufsichtigten Unternehmen, die Anforderungen an das IKT-Risikomanagement gemäß Art. 5 bis 15 DORA einzuhalten haben. Die Anforderungen an den vereinfachten IKT-Risikomanagementrahmen (Art. 16 DORA) bedürfen einer gesonderten Betrachtung. Sie sind nicht Gegenstand der vorliegenden Analyse.

Die Orientierungshilfe beruht u. a. auf Gesprächen mit Finanzunternehmen und stellt keine verbindliche DORA-Auslegung der BaFin dar.

Finanzunternehmen setzen KI entlang der gesamten Wertschöpfungskette ein. Dies geschieht aus unterschiedlichen Motiven: Ein wesentlicher Treiber für den Einsatz moderner KI-Technologie ist das Streben nach mehr Effizienz und optimierten Prozessen; ebenso erlauben komplexe mathematische Modelle eine präzisere Einschätzung von Risiken. Um die Bandbreite der Anwendung von KI bei Kreditinstituten und Versicherungsunternehmen zu illustrieren, bietet sich die Betrachtung von Beispielen aus den Wertschöpfungsketten dieser Unternehmen an.

Kreditinstitute nutzen KI im Vertrieb, um die Abwanderung von Kunden zu prognostizieren. In der Kreditvergabe können KI-Anwendungen die Sachbearbeitenden bei der Untersuchung von Jahresabschlussunterlagen unterstützen. Im Fondsmanagement wird KI genutzt, um große Mengen von Analystenberichten zu Anlageinstrumenten zusammenzufassen.

Bei Versicherungsunternehmen kommen interaktive KI-Assistenten (Chatbots) im Vertrieb und in der Kundenkommunikation zum Einsatz, um Mitarbeitende oder Kunden über Produkteigenschaften zu informieren. Das Pricing von Produkten kann mittels komplexer Modelle unter potentieller Verwendung von Echtzeitdaten präziser auf die versicherten Risiken abgestimmt werden (dynamisches Pricing bzw. Telematik). Im Underwriting können KI-Anwendungen bei der Einschätzung von Risiken unterstützen. Das automatisierte Inputmanagement erlaubt es, eine Vielzahl eingehender Dokumente effizient zu routen. Im Schadenmanagement nutzen Versicherer KI-Anwendungen, um die Schadensachbearbeiter in der Schadenregulierung zu unterstützen, bspw. bei der automatischen Auszahlung von Kleinschäden.

[1] Verordnung (EU) 2022/2554 über die digitale operationale Resilienz im Finanzsektor

[2] Mit Finanzunternehmen sind solche im Sinne der Definition in Art. 2 Abs. 2 Verordnung (EU) 2022/2554 (DORA) i. V. m. Art. 2 Abs. 1 a. – t DORA gemeint.

[3] Delegierte Verordnung (EU) 2024/1774

[4] Delegierte Verordnung (EU) 2025/532

Und in der Leistungsbearbeitung werden KI-Anwendungen zur Betrugserkennung eingesetzt. Derzeit dominieren die Anwendungen in den Bereichen, wo die größten Effizienzsteigerungen zu erzielen sind, also insb. in der Kundenkommunikation, im Schadenmanagement und in der Leistungsbearbeitung.

In Compliance-Funktionen wird KI genutzt, um regulatorische Entwicklungen zu überwachen. Die Unternehmenskommunikation kann KI einsetzen, um automatisiert Beiträge in sozialen Medien zu erstellen. Auch in Risikomodellen wird KI verwendet, um eine präzisere Ermittlung von Kapitalanforderungen zu erzielen.

Übergreifend ist der Einsatz von KI-Assistenten zu beobachten. Dabei handelt es sich zumeist um große Sprachmodelle, die unstrukturierte Daten wie etwa Texte und Bilder verarbeiten und erzeugen. Solche Assistenzsysteme können breit eingesetzt werden, um etwa Präsentationen, Programmcode oder Videos zu erstellen.

1. Begriffsbestimmung eines KI-Systems

Die Orientierungshilfe betrachtet die mit dem Einsatz von „KI-Systemen" verbundenen IKT-Risiken.

Der Begriff „KI-System" ist in Art. 3 Nr. 1 der EU KI-Verordnung[5] legaldefiniert. Demnach ist ein KI-System ein „maschinengestütztes System, das für einen in unterschiedlichem Grade autonomen Betrieb ausgelegt ist und das nach seiner Betriebsaufnahme anpassungsfähig sein kann und das aus den erhaltenen Eingaben für explizite oder implizite Ziele ableitet, wie Ausgaben wie etwa Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erstellt werden, die physische oder virtuelle Umgebungen beeinflussen können".

"Maschinengestützt" bezieht sich auf den Umstand, dass KI-Systeme mithilfe von Maschinen entwickelt und auf Maschinen betrieben werden. Die Maschine umfasst dabei Hardware und Softwarekomponenten, welche das Funktionieren des KI-Systems ermöglichen. Die Hardware-Komponenten beziehen sich auf die physischen Elemente der Maschine, wie z. B. die Verarbeitungseinheiten, Speicher, Netzwerkeinheiten, Ein- und Ausgabeschnittstellen. Die Softwarekomponenten umfassen u. a. den Quellcode, Betriebssysteme und Anwendungen (vgl. Leitlinie der Kommission vom 29.07.2025, C(2025) 5053 final, Rn. 11[6]).

„KI-Systeme" sind damit ein Unterfall der weiter gefassten „Netzwerk- und Informationssysteme" gemäß Art. 3 Nr. 2 Digital Operational Resilience Act (DORA). Aus dem Einsatz eines solchen maschinengestützten Systems und der darin genutzten Hard- und Software entstehen IKT-Risiken. Die Orientierungshilfe befasst sich ausschließlich mit diesen IKT-Risiken und legt die Anforderungen von DORA sowie der RTS RMF und RTS Unterauftragsvergabe zugrunde. Der Begriff „maschinengestütztes System" wird somit im Sinne von DORA, als eine Kombination von IKT-Assets (Hard- und Software) und -Infrastruktur verstanden, in die ein komplexes mathematisches Modell implementiert ist (vgl. auch Abbildung 1). Das Modell selbst wird hierbei als IKT-Asset (Software) verstanden.

Der in dieser Orientierungshilfe verwendete Begriff „KI-System" ist also als eine Kombination von IKT-Assets und IKT-Infrastruktur zu verstehen (vgl. auch Abbildung 1). Die in der Definition der KI-Verordnung enthaltenen Begrifflichkeiten zu Autonomie und Anpassungsfähigkeit sowie die mathematische Modellmethodik inklusive der verwendeten Daten, deren Entwicklung und Validierung werden nicht betrachtet.

KI-Systeme dienen – wie IKT-Systeme im Allgemeinen – den Geschäftsprozessen. Art und Umfang eines KI-Systems ist einzelfallabhängig und muss insbesondere dem Geschäftszweck, der Risikosituation sowie dem Sachzusammenhang entsprechen.

2. Gegenstand der Orientierungshilfe

Spezifische IKT-Risiken stehen in keinem Zusammenhang mit der Verortung von KI-Systemen in der Wertschöpfungskette eines Finanzunternehmens. Vielmehr ergeben sie sich aus der

[5] Verordnung (EU) 2024/1689 zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz

[6] Leitlinien der Kommission zur Definition eines Systems der Künstlichen Intelligenz gemäß der Verordnung (EU) 2024/1689 (KI-Verordnung)

Einbindung des KI-Systems in die IKT-Landschaft der Unternehmen. Daher betrachtet die Orientierungshilfe nicht die IKT-Risiken entlang der Wertschöpfungskette, sondern entlang des sogenannten KI-Lebenszyklus.

Entlang dieses Zyklus von der Datenbeschaffung über die Modellentwicklung und Bereitstellung bis hin zum laufenden Betrieb und der Stilllegung soll die Sicherheit und Resilienz eines KI-Systems gewährleistet werden. Neben spezifischen Schutzmaßnahmen für die IKT-Assets ist es entscheidend, dass KI-Systeme auch innerhalb des bestehenden IKT-Risikomanagementrahmens berücksichtigt werden.

Die Orientierungshilfe befasst sich ausschließlich mit IKT-Risiken und deren Behandlung unter DORA und ergänzenden technischen Regulierungsstandards.
Ein wichtiger Grundsatz von DORA ist der risikobasierte Ansatz und der Grundsatz der Verhältnismäßigkeit (Art. 4 DORA), der bei der Implementierung der Anforderungen Berücksichtigung findet. Somit benötigen insbesondere KI-Anwendungen, die in kritische oder wichtige Funktionen integriert sind, umfangreichere Sicherheits- und Kontrollmaßnahmen als z. B. KI-basierte Self-Service-Assistenten, die vollständig unter menschlicher Überwachung stehen und nicht in Entscheidungsprozesse eingebunden sind.

Diese Orientierungshilfe, die sich als nicht verpflichtende Hilfestellung versteht, definiert keine aufsichtlichen Erwartungen. Wie auch in DORA, sind stets der risikobasierte Ansatz und der Grundsatz der Verhältnismäßigkeit zu beachten.

3. Aufbau der Orientierungshilfe

Die Betrachtungen der Sicherheit von KI-Systemen werden folgendermaßen gegliedert:

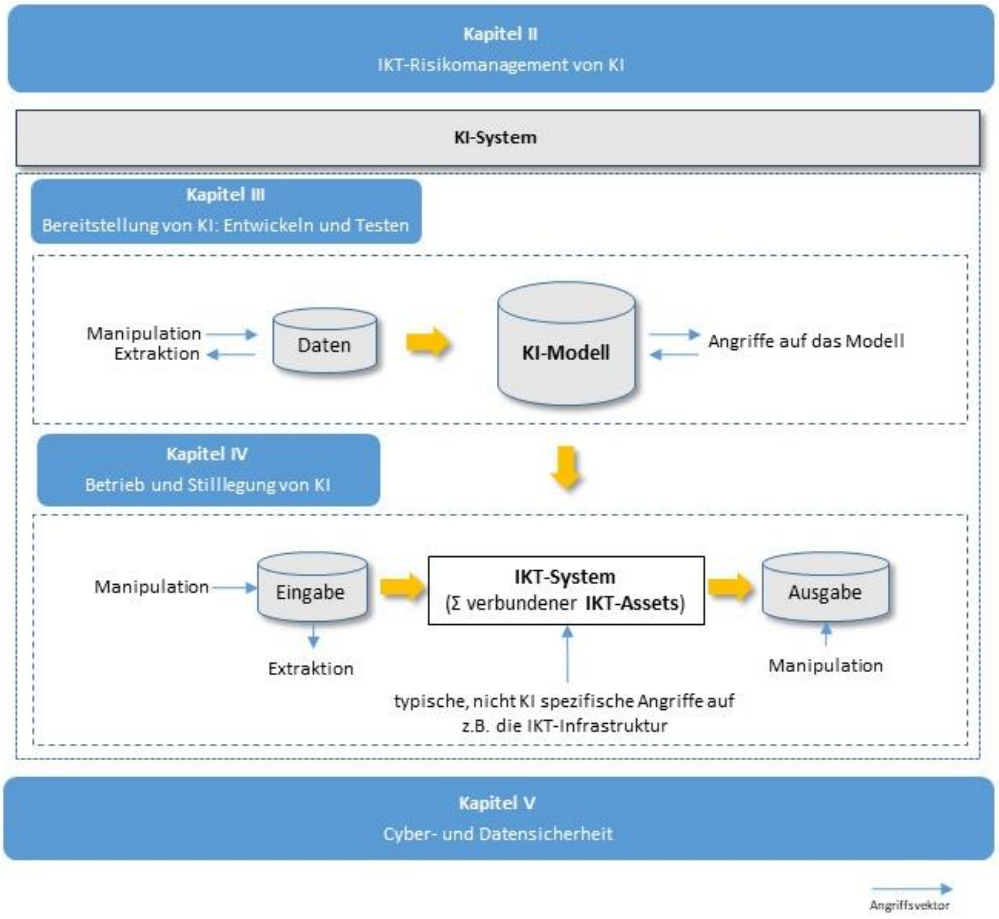
Zunächst erfolgt eine Untersuchung wesentlicher relevanter Aspekte für das IKT-Risikomanagement (Kapitel II), die lebenszyklusübergreifend Anwendung finden (vgl. Abbildung 1).

Es schließen sich Diskussionen der Anforderungen an KI-Systeme für das Entwickeln und Testen (Kapitel III), sowie für den Betrieb und Stilllegung (Kapitel IV) an. Enthalten sind hier die wesentlichen technischen Anforderungen zum Schutz des KI-Systems. Diese Kapitel berücksichtigen auch den Datenfluss und die für Manipulation, Datenextraktion und Angriffe gefährdeten IKT-Assets (vgl. Datenfluss und Angriffsvektoren in Abbildung 1).

Darauf folgt eine Betrachtung allgemein gültiger Vorgaben für KI-Systeme zur Cyber- und Datensicherheit (Kapitel V). Den Aspekten Cyber- und Datensicherheit kommt insofern eine Sonderrolle zu, da sie übergreifend für alle Elemente des KI-Lebenszyklus angewendet werden. Eine Schlussbetrachtung fasst die wichtigsten Erkenntnisse zusammen (Kapitel VI).

Die Orientierungshilfe ist als lebendes Dokument zu verstehen, das entsprechend des technischen Fortschritts und neuen regulatorischen Entwicklungen angepasst werden kann.

Abbildung 1: Gegenstand und Aufbau der Orientierungshilfe



II. IKT-Risikomanagement von KI

1. IKT-Risiken aus der Nutzung von KI

Die Implementierung und der Betrieb von KI-Systemen können erhebliche Risiken mit sich bringen. Aus regulatorischer Sicht besonders zu betrachten sind hierbei die IKT-Risiken von KI-Systemen. KI-Systeme werden analog zu allgemeinen IKT-Systemen hinsichtlich ihres Risikoprofils, ihrer Komplexität und der unterstützten Funktionen, etwa bei der Verarbeitung sensibler Daten oder der Unterstützung von kritischen oder wichtigen Funktionen, bewertet.

Zahlreiche KI-Systeme können ohne die Nutzung von Cloud-Dienstleistungen nicht betrieben werden. Daher kommt auch dem *IKT-Drittparteienrisiko* eine hohe Bedeutung zu. Hier spielt nicht nur die strategische Frage der Abhängigkeit von (wenigen) Anbietern eine Rolle.

Weiterhin kann die Sicherstellung der *Datenqualität und –integrität* eine große Herausforderung darstellen. Manipulierte oder fehlerhafte Daten können die Modellleistung und Sicherheit erheblich beeinträchtigen. Ebenso gefährdet die unkontrollierte Weiterverwendung oder unsichere Entsorgung von KI-Anwendungen sensible Modell- oder Unternehmensdaten.

Auch die *Cybersicherheit* erlangt durch die Nutzung komplexer KI-Systeme eine erhöhte Bedeutung. So können Schwachstellen, wie z. B. Backdoors während des Trainings unbemerkt in das Modell eingeschleust werden. Weiterhin kann die unsichere Bereitstellung von Software (etwa über Open-Source-Bibliotheken) dazu führen, dass Angreifer Modelle stehlen oder manipulieren. Ebenso können adversarielle Angriffe oder unautorisierte Zugriffe während der Nutzung des Modells zu Fehlentscheidungen der Unternehmen führen.

Diesen und weiteren IKT-Risiken zu begegnen, ist die Aufgabe des IKT-Risikomanagements. In den folgenden Abschnitten werden ausgewählte DORA-Anforderungen daran für KI-Systeme weiter detailliert.

2. Governance und Organisation

Das Risikomanagement beginnt auf strategischer Ebene mit der Implementierung einer geeigneten Governance- und Organisationsstruktur. Finanzunternehmen erstellen oftmals eine an der Gesamtstrategie, Risiko- und schließlich ggfs. IKT- sowie DOR-Strategie ausgerichtete KI-Strategie und lassen diese vom Leitungsorgan genehmigen. Bei der KI-Strategie kann es sich um eine eigenständige Strategie handeln, sie kann aber auch in eine übergeordnete Strategie integriert werden. Die Formulierung einer KI-Strategie gewinnt insbesondere dann an Gewicht, wenn KI-Anwendungen kritische oder wichtige Funktionen unterstützen.

KI-Strategie. Die Entwicklung einer klaren Technologie-Roadmap kann Grundlage der KI-Strategie sein, um die erforderlichen IKT-Ressourcen, -Kapazitäten und -Investitionen für den KI-Einsatz zu definieren. KI-Anwendungen können nahtlos in bestehende IKT-Systeme und -Prozesse integriert werden. Ein kontinuierliches Innovationsmanagement hilft, neue Technologien und Trends im Bereich KI zu identifizieren und zu bewerten.

Es bietet sich an, entlang eines Prozesses, alle für den Einsatz von KI relevanten Schritte von der Strategie über die Entwicklung bis hin zur Stilllegung abzudecken und zu dokumentieren. Dabei ist es wichtig, dass das beaufsichtigte Unternehmen zuerst alle für den Einsatz von KI relevanten Prozesse dahingehend überprüft, ob diese für Künstliche Intelligenz ausgelegt sind und ob eine Sensibilisierung zum Umgang mit Informationswerten besteht, bevor es ein KI-System implementiert.

Ein weiterer wichtiger Aspekt ist die Entwicklung der KI-Kompetenzen. Schulungen und Weiterbildungen stellen sicher, dass die Mitarbeitenden über die notwendigen Fähigkeiten und Kenntnisse im Umgang mit KI-Systemen verfügen. Sie gewährleisten somit den Erwerb von für den jeweiligen Aufgabenbereich angemessenen Kenntnissen in den Fachbereichen (Art. 13 Abs. 6 DORA). Die Förderung von Talenten und der Aufbau von Expertenteams im Bereich KI sind ebenso wichtig wie der interne Wissenstransfer durch Dokumentation und Schulungsveranstaltungen. Bei der Nutzung von KI-Systemen gewinnt zudem die interdisziplinäre Zusammenarbeit an Bedeutung, etwa zwischen der IT-Abteilung und den Fachabteilungen.

Governance/Organisation. Mit den Anforderungen an Governance und Organisation etabliert DORA einen zentralen Baustein zur Zielerreichung einer hohen digitalen Resilienz von Finanzunternehmen. Eine Schlüsselrolle kommt hier der Einrichtung eines internen Governance- und Kontrollrahmens zu. Dies soll ein wirksames und umsichtiges Management von IKT-Risiken gewährleisten.

Ein weiterer Baustein ist die hervorgehobene Rolle des Leitungsorgans. Neben der Gesamtverantwortung für die Festlegung und Genehmigung der Strategie für die digitale operationale Resilienz und die Zuweisung angemessener Budgetmittel für den IKT-Risikomanagementrahmen obliegt diesem Gremium, die Letztverantwortung für das Management der IKT-Risiken des Finanzunternehmens (Art. 5 Abs. 2 lit. a DORA).

Zudem müssen die Mitglieder des Leitungsorgans ausreichende Kenntnisse und Fähigkeiten u. a. durch spezielle Schulungen erwerben und auf dem neuesten Stand halten, um IKT-Risiken und deren Auswirkungen auf die Geschäftstätigkeit des Finanzunternehmens verstehen und bewerten zu können (Art. 5 Abs. 4 DORA). Darüber hinaus sind Verantwortlichkeiten innerhalb der Organisation und abhängig von der jeweils ausgeübten Funktion festzulegen, bspw. für die Verwendung von KI-generierten Ergebnissen in Entscheidungsprozessen.

In Bezug auf die allgemeinen Anforderungen ist es hilfreich, geeignete Vorgaben zu formulieren, die im Einklang mit der (DOR-)Strategie des beaufsichtigten Unternehmens und dessen Leit- und Richtlinien zur Informationssicherheit stehen.

Spezifische Regelungen für IKT-Drittdienstleister und deren Dienstleistungen sollten idealerweise u. a. eine Risikobewertung, Hinweise des Dienstleisters, sowie eigene Maßnahmen zur Risikoreduktion berücksichtigen. Dabei sollten die allgemeinen Anforderungen auf den Anwendungsfall bezogen werden. Und es sollte geprüft werden, ob zusätzliche KI-spezifische Maßnahmen notwendig sind. Insbesondere treffen viele Finanzunternehmen risikobasierte Vorgaben zur Nutzung des KI-Systems, die von der Kritikalität der verwendeten Daten und dem Ort der Speicherung und Verarbeitung abhängen.

Bei der Einführung von KI-Systemen ist es üblich, dass Governance-Rahmenwerke die Beteiligung der IKT-Risikomanagementfunktion, der Kontrollfunktionen und der internen Revisionsfunktionen vorsehen. Dies erfolgt in Abhängigkeit von der Kritikalität des KI-Systems und unter Wahrung der Unabhängigkeit und Vermeidung von Interessenskonflikten.

3. IKT-Risikomanagementrahmen nach DORA

Der Kern des IKT-Risikomanagements von KI-Systemen ist der sog. *IKT-Risikomanagementrahmen* (Art. 6 DORA). Dieser dient dazu, die IKT-Risiken, mit denen das Finanzunternehmen konfrontiert ist, effektiv zu managen.

> Artikel 6 Abs.1 DORA
>
> **IKT-Risikomanagementrahmen**
>
> Finanzunternehmen verfügen über einen soliden, umfassenden und gut dokumentierten IKT-Risikomanagementrahmen, der Teil ihres Gesamtrisikomanagementsystems ist und es ihnen ermöglicht, IKT-Risiken schnell, effizient und umfassend anzugehen und ein hohes Niveau an digitaler operationaler Resilienz zu gewährleisten.

In Kapitel II legt DORA über die Finanzdienstleistungssektoren hinweg harmonisierte und einheitliche Anforderungen an das IKT-Risikomanagement fest. Dies umfasst neben den bereits erläuterten Anforderungen an Governance und Organisation auch den IKT-Risikomanagementrahmen. Damit soll erreicht werden, dass die Finanzunternehmen eine für sie angemessene digitale operationale Resilienz erzielen. Sie sollen also so widerstands- und anpassungsfähig werden, dass sie ihre digitalen operationellen Prozesse auch während und nach einem IKT bezogenen Vorfall aufrechterhalten können.

Als Bestandteil des Gesamtrisikomanagements umfasst der IKT-Risikomanagementrahmen unter anderem folgende Anforderungen:

- Identifizierung gem. Art. 8 DORA,
- Schutz und Prävention gem. Art. 9 DORA,
- Erkennung gem. Art. 10 DORA,
- Reaktion und Wiederherstellung gem. Art. 11 DORA,
- Lernprozesse und Weiterentwicklung gem. Art. 13 DORA sowie
- Kommunikation gem. Art. 14 DORA.

Analog zu anderen IKT-Assets, sind auch KI-Systeme in den IKT-Risikomanagementrahmen zu integrieren. Dadurch werden die Ermittlung von Schwachstellen (etwa im Modelltraining, in Datenpipelines oder bei der Inferenz) sowie die Bewertung quantitativer und qualitativer Risikokriterien umfasst (Art. 8 DORA). Weiterhin müssen Maßnahmen zur Behandlung identifizierter Risiken wie etwa der Einsatz adversarialer Trainingsmethoden oder die Überwachung von Modelldrift dokumentiert und regelmäßig überprüft werden (Art. 9 DORA).

Der IKT-Risikomanagementrahmen muss mindestens jährlich überprüft werden (Art. 6 Abs. 5 Satz 1 und 2 DORA). Weiterhin müssen Finanzunternehmen auf Anfrage der zuständigen Behörde einen Bericht über die Überprüfung ihres IKT-Risikomanagementrahmens in einem durchsuchbaren elektronischen Format vorlegen können. Der Bericht muss u. a. den aktuellen

Risikostatus, die ergriffenen Maßnahmen und identifizierte Schwächen dokumentieren (Art. 27 RTS RMF). Dieser Bericht kann bei Bedarf auch um spezifische Informationen zu KI-Systemen angereichert werden.

III. Bereitstellung von KI: Entwickeln und Testen

1. Softwareentwicklung

Wenn Finanzunternehmen KI-Systeme selbst entwickeln, dann geschieht dies in der Regel innerhalb der dafür vorgesehenen Regelprozesse. So haben sie die volle Kontrolle über die Software und die mathematischen Modelle. Die Entwicklung von Software erfordert nicht nur ausreichend spezialisierte Kenntnisse und Fähigkeiten, sondern stellt auch Anforderungen an die Gestaltung des Entwicklungsprozesses. Auch hier finden die Vorgaben von DORA für KI-Systeme Anwendung.

Kenntnisse und Fähigkeiten. Die Entwicklung von KI-Systemen stellt hohe Anforderungen an die Fähigkeiten der Beschäftigten und der Geschäftsleitung. Beschäftigte, die mit Aufgaben im Zusammenhang mit KI betraut sind, sehen sich der Herausforderung gegenüber, angemessene und einschlägige Kompetenzen und Kenntnisse über die Funktionsweise der relevanten KI-Systeme, der mit ihnen verbundenen Risiken sowie der mit dem Cloud- und On-Premise-Betrieb verbundenen technischen und organisatorischen Besonderheiten zu erwerben.

Gerade die Möglichkeit, Software mithilfe von KI-Assistenten zu entwickeln, kann Fachbereiche außerhalb der IKT-Funktion in die Lage versetzen, auch komplexe Anwendungen selbst zu erstellen. Hier gelten ebenfalls die zuvor genannten Anforderungen an Kenntnisse und Fähigkeiten (Art. 16 Abs. 9 RTS RMF), auch in Hinblick auf das Testen.

Auch das Management und insbesondere die Mitglieder des Leitungsorgans können durch ein ausreichend tiefes Verständnis von KI-Systemen dazu beitragen, die Risiken aus der Softwareentwicklung einzuschätzen und zu managen (Art. 5 Abs. 4 und Art. 13 Abs. 6 DORA).

Der Umfang der notwendigen Kenntnisse hängt von den jeweiligen Aufgaben ab: Je technischer die Aufgaben, desto spezifischer das Wissen. Aufgrund des rasanten technischen Fortschritts im Bereich der Künstlichen Intelligenz sind regelmäßige Schulungen der Mitarbeitenden und eine kontinuierliche Beobachtung der Entwicklungsdynamik erforderlich. Typischerweise werden Schulungsmaßnahmen für Mitarbeitende definiert, die an der Entwicklung, dem Betrieb und der Wartung von KI-Systemen beteiligt sind (Art. 13 Abs. 6 DORA).

Entwicklungsprozess. Bei der Einführung und Weiterentwicklung von KI-Systemen hilft die Etablierung eines robusten Projektmanagements, das alle Phasen (Planung, Entwicklung, Test, Rollout und Betrieb) abdeckt. Hierzu gehören klare Zielsetzungen, die Festlegung von Governance-Strukturen, detaillierte Projektrisikobewertungen sowie Testverfahren (s. u.) (Art. 15 RTS RMF).

Bei der Entwicklung und Validierung von KI-Systemen haben sich gängige Praktiken des Software-Engineerings wie Unit-Tests, Integrationstests und Code-Reviews bewährt. Regelmäßige, bestenfalls automatisierte Test- und Validierungsverfahren eignen sich, um Genauigkeit

und Zuverlässigkeit sicherzustellen. Eine sichere und isolierte Umgebung für die experimentelle Entwicklung und das Testen von KI-Systemen ist essenziell.

Eine umfassende Dokumentation in Form von technischen Spezifikationen kann zur Nachvollziehbarkeit und Transparenz der Entwicklung von KI-Systemen beitragen. Neben den technischen IKT-Spezifikationen ist auch die Beschreibung der verwendeten Algorithmen, Daten und Parameter sinnvoll (Art. 16 RTS RMF).

KI-Systeme werden analog zu IKT-Systemen gemäß den Sicherheits- und Funktionalitätsanforderungen beschafft, entwickelt und gewartet. Dabei werden typischerweise technische Spezifikationen definiert, die auch Anforderungen an die Sicherheit enthalten, beispielsweise den Schutz vor Manipulationen (Art. 16 RTS RMF).

Jede Änderung an KI-Systemen (sei es an der Software oder Hardware) unterliegt typischerweise einem strengen Änderungsmanagement, das eine unabhängige Überprüfung, dokumentierte Tests und die Angabe von Ausweichverfahren beinhaltet. Zudem sind Verfahren, Protokolle und Tools für den Umgang mit Notfalländerungen zu etablieren, die angemessene Schutzvorkehrungen vorsehen. Das Ziel besteht darin, das Risiko unbeabsichtigter Sicherheitslücken oder Manipulationen zu minimieren (Art. 17 RTS RMF).

Versionskontrollsysteme sind sinnvoll, um Änderungen an KI-Modellen und -Anwendungen nachzuverfolgen. Dies ermöglicht das Zurückverfolgen von Fehlern und die Reproduktion von Ergebnissen. Eine systematische Archivierung aller Modellversionen und -parameter stellt die Wiederverwendbarkeit und Nachvollziehbarkeit sicher (Art. 17 Abs. 1 bis 2 RTS RMF). DORA trifft keine Unterscheidung zwischen Anwendungen, die innerhalb der IKT-Funktion und außerhalb der IKT-Funktion (oftmals als „individuelle Datenverarbeitung" (IDV) bezeichnet) entwickelt werden (Art. 16 Abs. 9 RTS RMF). Falls eine Entwicklung von KI-Systemen außerhalb der IKT-Funktion zwingend erforderlich ist, erfolgt diese daher auch entlang der zuvor beschriebenen Prozesse.

In der Praxis haben sich mehrere Vorgehensweisen als hilfreich bei der Entwicklung und Änderung von KI-Systemen herausgestellt, um die Anforderungen von DORA zu erfüllen. Für den Trainingsprozess schließen diese Maßnahmen ein:

- Sicherstellung, dass Trainings- und Testdaten aus vertrauenswürdigen Quellen stammen.
- Sicherstellung, dass (Open-Source-)Bibliotheken und Softwarepakete, die für das Training genutzt werden, keine bekannten Schwachstellen enthalten.
- Dokumentation und Versionierung des gesamten Trainingsprozesses, um Transparenz und Nachvollziehbarkeit sicherzustellen.
- Einsatz von Sicherheitsanalysen, um versteckte Manipulationen im Modell oder den Trainingsdaten zu identifizieren

Nutzung von Open-Source-Bibliotheken. Selbstentwickelte sowie zugekaufte Software für KI-Systeme kann auch Open-Source-Bibliotheken nutzen. Im äußersten Fall kann dies auch die Nutzung eines Open-Source-Modells umfassen, das auf der Hardware eines Finanzunternehmens selbst implementiert, trainiert und betrieben wird. Die bereits geschilderten Aspekte des Entwicklungsprozesses haben sich auch hier als relevant erwiesen. Bei Open-Source-Software besteht zusätzlich das Risiko, dass Schadcode über diese Bibliotheken eingeschleust wird. Darüber hinaus besteht die Gefahr, dass die Bibliotheken nach einigen Jahren nicht mehr gepflegt werden. In diesem Fall kann das KI-System identifizierte, aber nicht behobene Schwachstellen enthalten und nicht mehr dem neuesten Entwicklungsstand entsprechen.

Code-Erzeugung mit KI-Assistenten. Softwareentwickler nutzen zunehmend KI-Assistenten, um Anwendungen effizienter zu erstellen. Grundsätzlich gelten für alle Arten der Codeerstellung, sei es durch Menschen oder Maschinen, die gleichen Regelungen. Eine Herausforderung besteht dennoch darin, den maschinell erzeugten Code darauf zu prüfen, inwieweit dem Nutzer nicht bekannte, KI-basierte Funktionen aufgerufen werden. Dies kann u. a. durch eine statische Codeanalyse (Art. 16 Abs. 3 RTS RMF) ermittelt werden. Tiefere Quellcode-Analysen und -Prüfungen der Qualität sowie der Nachvollziehbarkeit und Verständlichkeit können für zusätzliche Sicherheit sorgen. Für das nachfolgende Testen hat sich eine angemessene Dokumentation des Quellcodes als empfehlenswert erwiesen.

2. Testen von KI

Ein wesentlicher Schritt bei der Bereitstellung von KI-Systemen ist das Testen von Hard- und Software. Auch hierfür sind die allgemeinen Vorgaben von DORA entsprechend auf KI-Systeme anzuwenden.

Der RTS RMF beschreibt detailliert die zum Testen von IKT-Systemen durchzuführenden Maßnahmen (Art. 16 Abs. 2 RTS RMF). So sind beispielsweise Sicherheitstests im Rahmen der Quellcodeüberprüfung für internetfähige Systeme und Anwendungen notwendig (Art. 16 Abs. 3 RTS RMF).

Testumfang. Finanzunternehmen müssen Tests entwickeln, dokumentieren und implementieren. Dabei muss der Testumfang der Kritikalität der betreffenden Geschäftsprozesse und IKT-Assets angemessen sein. Die Tests müssen so ausgelegt sein, dass überprüft werden kann, ob neue IKT-Systeme – und somit auch KI-Systeme – ihrer geplanten Bestimmung angemessen sind. Dies schließt auch die Qualität der intern entwickelten Software mit ein (Art. 16 Abs. 2 RTS RMF).

Quellcode. Der RTS RMF enthält spezifische Anforderungen an den Umgang mit Quellcode aus der Anwendungsentwicklung. So muss der Quellcode vor dem produktiven Einsatz unter Einsatz von statischen und dynamischen Testverfahren auf Anomalien überprüft werden (Art. 16 Abs. 3 RTS RMF).

Insbesondere bei der Einbindung von Open-Source-Bibliotheken ist es sinnvoll, sicherzustellen, dass die benutzten Funktionen keine KI-basierten Funktionalitäten beinhalten, deren Risiken dem Finanzunternehmen nicht bekannt sind oder die nicht mitigiert werden können (Art. 16 Abs. 3 RTS RMF).

Der Test beinhaltet zudem, dass proprietäre Software (kompilierter Quellcode) sowie nach Möglichkeit der Quellcode, der von IKT-Drittdienstleistern bereitgestellt wird oder aus Open-Source-Projekten stammt, vor der Inbetriebnahme analysiert und getestet wird (Art. 16 Abs. 8 RTS RMF).

Eine Herausforderung bei den Tests ist es, zu identifizieren, ob Anwendungen externe KI-Modelle etwa über APIs (Anwendungsprogrammierschnittstellen) einbinden. Dadurch kann eine ursprünglich als Nicht-KI-Anwendung geplante Software zu einem KI-System werden.

Generative KI. Besondere Herausforderungen für das Testen ergeben sich bei der Nutzung generativer KI. Die typischerweise verwendeten großen Sprachmodelle (Large Language Models, LLMs) verfügen über eine komplexe interne Struktur (sog. Transformer-Struktur) mit einer Vielzahl peripherer Modelle, die an das eigentliche Modell angebunden sind. Die Transformer-Struktur basiert auf tiefen neuronalen Netzen mit mehreren Milliarden Parametern (sogenannten Gewichten). Das Training von LLMs wird auf Basis großer Datensätze (z. B. Wikipedia) durch den Anbieter durchgeführt. Der Nutzer kann das Modell nachtrainieren (sog. „Finetuning"), was jedoch mit einem hohen Rechenaufwand verbunden ist.

Da generative KI für allgemeine Verwendungszwecke geeignet ist, gestaltet sich das Testen schwieriger als bei selbst entwickelter Software, die nur für einen bestimmten Anwendungszweck entwickelt wurde. Grundsätzliche Testverfahren für LLMs können die interne Struktur der Modelle berücksichtigen (z. B. durch Betrachtung der Wahrscheinlichkeiten für die Erzeugung eines bestimmten Token-Streams bei gegebenem Eingabe-Prompt) oder agnostisch ausgestaltet sein. Im letzteren Fall beantwortet das Modell beispielsweise Testfragen, und die Antworten werden durch einen Menschen oder ein weiteres Modell auf ihre Qualität hin beurteilt. Eine weitere Herausforderung für das Testen ergibt sich durch unangekündigte Modelländerungen bei der Nutzung eines KI-Modells, das von Dritten bezogen wurde.

In der Entwicklung und Änderung von KI-Systemen haben sich mehrere Vorgehensweisen als geeignet erwiesen, um die Schutzziele von DORA zu erfüllen. Je nach Kritikalität können folgende Maßnahmen sinnvoll sein:

- Simulation von Angriffen auf KI-Systeme (Adversarial Testing, z. B. Data Poisoning, Evasion Attacks).
- Durchführung von Adversarial Penetration Tests (z. B. in Abstimmung mit dem Cloud-Anbieter), um KI-spezifische Angriffe zu simulieren.
- Überprüfung der Performance von KI-Systemen in Stresstests, z. B. mit stark veränderten Datenverteilungen oder in Überlastszenarien.
- Bei nicht selbst entwickelten KI-Systemen kann es auch in Betracht kommen, den Hersteller dieses Systems in die Tests einzubeziehen und ggf. entsprechende Nachweise über deren Ausführung zu erhalten.

IV. Betrieb und Stilllegung von KI

1. Prozesse für den Betrieb und die Deinstallation von KI-Systemen

Für den Betrieb von IKT-Systemen müssen entsprechende Prozesse definiert werden. Diese Prozesse berücksichtigen idealerweise die Spezifika der KI-Systeme, z. B. ob es sich um die Anbindung eines LLM von einem Cloud-Dienstleister handelt oder um selbstentwickelte Software im eigenen Rechenzentrum.

Bei der Definition von Betriebsprozessen ist es notwendig, den gesamten Lebenszyklus von KI-Systemen abzudecken. Dies umfasst u. a. Anforderungen an die Entwicklung, Installation und Betrieb und Deinstallation von KI-Systemen sowie die Dokumentation aller Betriebsaktivitäten wie z. B. Logfiles von Modellinferenzprozessen (Art. 8 RTS RMF).

Identifikation und Dokumentation. Für das Management von Informations- und IKT-Assets fordert DORA eine Richtlinie und Verfahren (Art. 8 Abs. 4 DORA i. V. m. Art. 4 und 5 RTS RMF). Informations- bzw. IKT-Assets (inkl. Komponenten von KI-Systemen) sind zu ermitteln, zu klassifizieren und zu dokumentieren sowie kontinuierlich zu überwachen. Dies betrifft u. a. auch Trainingsdatensätze, Implementierungen von Modellen, Softwarebibliotheken, Hardware, selbst und fremd erstellte Software. Klare Kriterien zur Kritikalität dieser Assets unterstützen zusätzlich bei der Klassifizierung.

Die Dokumentation von Informations- und IKT-Assets führt insbesondere dazu, dass eindeutig nachvollzogen werden kann, woher z. B. Trainingsdaten stammen, an welcher Stelle innerhalb des Lebenszyklus diese gespeichert werden und welche Zuständigkeiten und Verantwortlichkeiten hierfür bestehen (Art. 4 Abs. 2 lit. b RTS RMF).

Kapazitäten und Leistung. Es ist sinnvoll, auch KI-Systeme hinsichtlich ihrer Ressourcenanforderungen und Leistungsfähigkeit regelmäßig zu überprüfen, um Kapazitätsengpässe zu vermeiden und eine stabile Verfügbarkeit zu gewährleisten. Automatisierte Überwachungsverfahren sind einzusetzen, um die Effizienz und die Skalierbarkeit der KI-Infrastruktur sicherzustellen (Art. 9 RTS RMF).

Deinstallation. Es ist zielführend, in Richtlinien und Verfahren des Finanzunternehmens auch Vorgaben für die Deinstallation von KI-Systemen aufzunehmen, KI-Modelle nach dem Löschen unwiederbringlich zu entfernen und die Deaktivierung von veralteten Modellversionen zu regeln, um Missbrauch zu vermeiden (Art. 8 Abs. 2 lit. a i) RTS RMF).

Angriffe. KI-Systeme müssen während des Betriebs gegen gängige Cyberbedrohungen wie Adversarial Attacks, Model Poisoning oder Inference Attacks geschützt werden. Um diesen Bedrohungen zu begegnen, bedarf es geeigneter technische Sicherheitsmaßnahmen (Art. 9 Abs. 3 DORA). Dies wird in der Fallstudie zum Einsatz von KI-Assistenten vertieft.

Zudem sind rollenbasierte Zugriffsrechte für KI-Modelle und Trainingsdaten ein wirksames Werkzeug. Diese sind regelmäßig zu prüfen und zu dokumentieren (Art. 21 RTS RMF).

Artikel 10 DORA

Erkennung

Finanzunternehmen verfügen über Mechanismen, um anomale Aktivitäten (u. a. Probleme bei der Leistung von IKT-Netzwerken und IKT-bezogene Vorfälle), umgehend zu erkennen und potenzielle einzelne wesentliche Schwachstellen zu ermitteln.

Erkennung. Zur Früherkennung von Anomalien (Art. 10 DORA) hat es sich als sinnvoll erwiesen, KI-Systeme kontinuierlich zu überwachen. Dies ermöglicht es, Abweichungen vom erwarteten Verhalten frühzeitig zu erkennen.

Ein Bestandteil der Früherkennung sind regelmäßige, automatisierte Schwachstellenscans und -bewertungen. Dies schließt die Überprüfung von verwendeten Softwarebibliotheken, Frameworks und dem Quellcode von selbst entwickelter oder von Dritten bezogener KI-Software ein. Auch die Festlegung von klaren Fristen und Eskalationsprozessen für die Implementierung von Patches ist hilfreich (Art. 10 RTS RMF).

Weiterhin empfiehlt es sich, risikobasiert und in Abhängigkeit vor der Kritikalität sowie dem Einsatzgebiet der jeweiligen KI-Systeme eine lückenlose Protokollierung von KI-Entscheidungen, Modellversionen und Trainingsdaten (soweit datenschutzrechtlich zulässig) vorzusehen. Für KI-Systeme, die kritische oder wichtige Funktionen unterstützen, ist die Definition von Schwellenwerten und Indikatoren für Fehlverhalten vorteilhaft. Zudem wird empfohlen, sie regelmäßig auf ihre Wirksamkeit hin zu überprüfen (Art. 10 Abs. 1 UAbs. 2 i. V. m. Abs. 2 DORA).

Artikel 11 DORA

Reaktion und Wiederherstellung

Ein wirksames und angemessenes IKT-Geschäftsfortführungsmanagement begrenzt die Auswirkungen von IKT-bezogenen Vorfällen und stellt die Fortführung von kritischen oder wichtigen Funktionen im Finanzunternehmen sicher.

Reaktion und Wiederherstellung sowie Backup. Der Ausfall von KI-Systemen, bspw. durch IKT-bezogene Vorfälle, kann zu schwerwiegenden Folgen für Finanzunternehmen führen.

In Abhängigkeit von ihrer Kritikalität ist es sinnvoll, KI-Systeme entsprechend im IKT-Geschäftsfortführungsmanagement und dabei insb. in den o.g. Plänen zu berücksichtigen. Dabei sind auch Wiederherstellungszeiten und -punkte (Art. 12 Abs. 6 DORA) einzubeziehen.

Um ihre Funktionsfähigkeit im Ernstfall sicherzustellen, müssen regelmäßige Tests (mindestens jährlich) dieser Pläne durchgeführt und dokumentiert werden (Art. 11 Abs. 6 lit. a DORA und Art. 25 RTS RMF).

Finanzunternehmen müssen auch IKT-Drittdienstleister, von denen sie KI-Systeme beziehen entsprechend in das IKT-Geschäftsfortführungsmanagement einbeziehen (Art. 11 Abs. 4 und Art. 28 DORA).

Zudem sollen Verfahren und Methoden für Wiedergewinnung und Wiederherstellung als auch Richtlinien und Verfahren zur Sicherung von KI-Systemen und deren Daten (z. B. regelmäßige Backups von Modellartefakten und Datensätzen), berücksichtigt werden

(Art. 12 Abs. 1 bis 3 und 7 DORA). Ihr Umfang richtet sich im Wesentlichen nach der Kritikalität der Informationen und der Vertraulichkeit der Daten. Die Sicherheit der Netzwerk- und Informationssysteme als auch die Schutzziele der Daten dürfen dabei nicht gefährdet werden. Bei einer Nutzung von durch Drittanbieter bereitgestellten KI-Systemen (z. B. über einen API-Zugriff) wird diese Aufgabe in der Regel vom Anbieter übernommen. Spezifika für durch Dritte bereitgestellte KI-Systeme finden sich im nächsten Abschnitt wieder.

Weiter ist es sinnvoll, adäquate Redundanzen für KI-Systeme im Einklang mit dem jeweiligen Geschäftsbedarf zu unterhalten (Art. 12 Abs. 4 DORA).

Erkenntnisse aus IKT-bezogenen Vorfällen sowie aus Herausforderungen bei der Aktivierung von IKT-Geschäftsfortführungsplänen und IKT-Reaktions- und Wiederherstellungsplänen sind systematisch auszuwerten und in die Weiterentwicklung der Systeme, Modelle und Prozesse zu integrieren (Art. 13 Abs. 3 DORA). Dies betrifft auch den IKT-Risikobewertungsprozess von KI-Systemen.

2. Cloud-Spezifika beim Betrieb von KI

In der am 1. Februar 2024 veröffentlichten Aufsichtsmitteilung zu Auslagerungen an Cloud-Anbieter teilten BaFin und Deutsche Bundesbank ihre gemeinsame Einschätzung zu Auslagerungen an Cloud-Anbieter mit. Einige dieser Aspekte können aufgrund des Betriebs von KI in der Cloud auch auf KI-Systeme übertragen werden. Im Folgenden werden Überschneidungen und ausgewählte KI-Spezifika aufgezeigt.

Risikobewertung und Due Diligence (Cloud-Aufsichtsmitteilung, Kap. III.2). Vor Abschluss eines Vertrages mit einem Cloud-Anbieter sollen Finanzunternehmen eine umfassende Risikobewertung (Art. 28 Abs. 4 lit. c DORA) durchführen, die sowohl technische als auch operative und regulatorische Aspekte berücksichtigt. Hierbei sollten auch Änderungen am KI-System berücksichtigt werden (z. B. Neutraining oder Anpassung der Modellstruktur), auch wenn alle Änderungen am Modell durch den Dienstleister erfolgen. Vor Vertragsabschluss mit einem Cloud-Anbieter ist eine Risikoinventur der KI-Anwendung (Wesentlichkeit, Sensibilität der Daten, technische Anforderungen) anzuraten. Zudem ist es zielführend, den gebotenen Sorgfaltspflichten (Due Diligence, Art. 28 Abs. 4 lit. d DORA) nachzukommen und die Eignung des Cloud-Anbieters sicherzustellen. Die erstreckt sich auch auf mögliche Interessenskonflikte (Art. 28 Abs. 4 lit. e DORA). Dabei sollte das Finanzunternehmen bei der Nutzung von Cloud-Dienstleistungen auch die Risiken eines unautorisierten Datenabflusses – auch an den Cloud-Anbieter –berücksichtigen.

Cyber- und Informationssicherheit (Cloud-Aufsichtsmitteilung, Kap. IV.2). Im Rahmen der Cyber-Sicherheit sollen potentielle Risiken aus Angriffsvektoren (u. a. Adversarial Attacks, Datenpoisoning) bei extern gehosteten Trainings- oder Inferenzumgebungen identifiziert werden. Weiterhin sollen IKT-Drittdienstleister (z. B. Cloud-Anbieter, externe Data-Science-Services) vor Vertragsabschluss hinsichtlich Sicherheitsstandards, Zertifizierungen und Datenschutz bewertet werden. Dabei sollen diese IKT-Drittdienstleister zumindest angemessene Standards für die Informationssicherheit einhalten (Art. 28 Abs. 5 DORA). Sollen kritische oder wichtige Funktionen unterstützt werden, sind bei der Auswahl die aktuellsten und höchsten Qualitätsstandards für die Informationssicherheit zu berücksichtigen (Art. 28 Abs. 5 DORA).

Artikel 28 DORA

Allgemeine Prinzipien des Managements des IKT-Drittparteienrisikos

Finanzunternehmen managen das IKT-Drittparteienrisiko als integralen Bestandteil des IKT-Risikos innerhalb ihres IKT-Risikomanagementrahmens.

Verträge und Unterauftragsvergabe (Cloud-Aufsichtsmitteilung, Kap. III.5). Es ist klar zu regeln, ob und unter welchen Bedingungen der Cloud-Anbieter weitere Unterauftragnehmer (Weiterverlagerung, Subdelegation, bzw. Untervergabe) einsetzen darf (Art. 30 Abs. 2 lit. a DORA). Bei KI-spezifischen Unterauftragsvergaben die kritische oder wichtige Funktionen unterstützen (z. B. spezialisierte ML-Libraries, GPU-Farmen, KI-Sicherheitsservices) überblickt das Institut jederzeit, welche Parteien an der Datenverarbeitung beteiligt sind (Art. 3 Abs. 6 ITS Informationsregister), wo (Standort/Region) diese stattfindet (Art. 30 Abs. 2 lit b DORA) und wie sich potenziell lange oder komplexe Ketten der Unterauftragsvergabe auswirken können (Art. 29 Abs. 2 DORA).

Finanzunternehmen müssen bei KI-Anwendungen, die kritische oder wichtige Funktionen unterstützen, klare Service Level Agreements (SLAs) und Sicherheitsvereinbarungen mit IKT-Drittparteien abschließen (Art. 30 Abs. 3 lit. a und lit. c DORA). In den SLAs soll klar und eindeutig verständlich definiert werden, wie Performance, Verfügbarkeit und Sicherheitsanforderungen eingehalten werden. Da KI-Systeme besonders empfindlich auf Latenz und Rechenkapazitäten reagieren, decken diese SLAs typischerweise auch diese Aspekte mit ab.

Weiterhin sollten Audit- und Kontrollrechte lückenlos auch bei Unterauftragsvergaben gelten, die kritische oder wichtige Funktionen unterstützen (Art. 3 Abs. 1 lit. d und Art. 4 Abs.1 lit. j RTS Untervergabe), um z. B. im Prüfungsfall Einblick in relevante KI-Logs, Trainingsumgebungen und Sicherheitsmaßnahmen zu erhalten. Zudem soll eine Überprüfung stattfinden, ob der Anbieter die regulatorischen Erwartungen (u. a. Cloud-Aufsichtsmitteilung, SLA-Anforderungen, Unterauftragnehmer-Transparenz) erfüllt (Art. 3 Abs. 1 lit. c und j RTS Untervergabe).

Exit-Strategie und Umzugsfähigkeit (Cloud-Aufsichtsmitteilung, Kap. IV.4). Ein wesentlicher Aspekt der Cloud-Nutzung ist die Beendigung bzw. der Wechsel von Cloud-Auslagerungen. Finanzunternehmen sollen im Rahmen einer Exit-Strategie für KI-Anwendungen, die kritische oder wichtige Funktionen unterstützen, sicherstellen, dass sie bei Problemen, wie etwa bei Non-Compliance mit aufsichtlichen Anforderungen des Cloud-Dienstleisters oder bei einem Wechsel des Anbieters, den Betrieb aufrechterhalten bzw. zeitnah migrieren können (Art. 28 Abs. 7 und Abs. 8 DORA). Dabei empfiehlt es sich, auf einen problemlosen Export der Modelle, Trainingsdaten und Konfigurationsskripte zu achten, um die Portierung in eine andere Umgebung zu gewährleisten.

Finanzunternehmen sollen sich darüber bewusst sein, dass proprietäre Cloud-Funktionen (z. B. spezielle KI-Angebote) die Kosten, die Dauer und notwendige technische Expertise bei einem Wechsel deutlich erhöhen können (sog. Vendor-Lock-In). Sie sollen daher in der Risikobewertung explizit berücksichtigt werden. Vor Vertragsabschluss ist zu klären, in welchen Formaten Trainingsdaten, Modelle und Metadaten exportiert werden können (z. B. Docker-Images, Speichersnapshots), um bei Ausfall oder Wechsel des Cloud-Anbieters weiterhin handlungsfähig zu bleiben (Art. 28 Abs. 8 und Art. 30 Abs. 3 lit. f DORA). Dabei sollen diese

Daten periodisch auf einem von diesem Cloud-Anbieter unabhängigen Speicherort abgelegt werden.

Tests von Szenarien für den Ernstfall (z. B. Cloud-Ausfall, Datenkorruption, Probleme mit einem Unterauftragnehmer) sind bei KI-Anwendungen die kritische oder wichtige Funktionen unterstützen regelmäßig durchführen, siehe Art. 28 Abs. 8 UAbs. 3 DORA.

Prüf- und Kontrollrechte (Art. 30 Abs. 3 lit. e DORA, Cloud-Aufsichtsmitteilung, Kap. III.5.3). Für wesentliche Auslagerungen oder vertragliche Vereinbarungen über die Nutzung von IKT-Dienstleistungen zur Unterstützung kritischer oder wichtiger Funktionen muss der Aufsicht (insb. BaFin) und dem auslagernden Unternehmen das Recht auf Prüfung beim Cloud-Anbieter eingeräumt werden.

V. Cyber- und Datensicherheit

KI-Systeme können als Einfallstore für Cyber-Angriffe und Datenmanipulation genutzt werden. Für alle Elemente des KI-Lebenszyklus sollten daher Aspekte der Cyber- und Datensicherheit beachtet werden. Nur so kann ein sicherer und resilienter Betrieb von KI insb. im Rahmen des IKT-Risikomanagementrahmens nach DORA gewährleistet werden. Die folgenden Hinweise enthalten Aspekte, die im Gegensatz zu Entwicklung und Betrieb lebenszyklusübergreifend gelten und KI-spezifische Besonderheiten beinhalten.

1. Cybersicherheit

KI-Systeme sind attraktive Ziele für Cyberangriffe, insbesondere, weil sie sensible Daten verarbeiten und ggf. in Entscheidungsprozesse eingebunden sind (vgl. Angriffsvektoren in Abbildung 1). Umfangreiche Sicherheitsmaßnahmen und regelmäßige Sicherheitsüberprüfungen sind daher von größter Bedeutung.

DORA fordert, dass Finanzunternehmen IKT-Sicherheitsrichtlinien entwickeln (Art. 9 Abs. 2 DORA). In diesen sollen auch KI-Systeme angemessen Berücksichtigung finden. Die Richtlinien sollen dabei insbesondere Maßnahmen zur Netzwerksicherheit, zur Datenübermittlung (Art. 2 Abs. 1 RTS RMF) und zum Schutz vor Datenmissbrauch umfassen. Für KI-Systeme bedeutet dies, dass neben der Absicherung der Modellzugänge auch der sichere Datenaustausch zwischen Trainingsdaten, Modellkomponenten und Endanwendern gewährleistet sein soll.

Systemsicherheit. Eine robuste Daten- und Systemsicherheit (Art. 11 RMF RTS) ist essenziell, um Cyberangriffe abzuwehren. Proaktive Abwehrmechanismen, die auf Bedrohungsanalysen und aktuellen Sicherheitsinformationen basieren, sind notwendig.

Diesbezüglich haben sich einige Maßnahmen als geeignet herausgebildet:

- Nutzung von Firewalls, IDS/IPS-Systemen und Zero-Trust-Modellen, um Angriffe auf die KI-Infrastruktur zu verhindern.
- Einsatz von Data Loss Prevention (DLP)-Mechanismen, um das unerlaubte Abgreifen von KI-Daten zu verhindern.
- Spezialisierte Resilienztests für KI-Systeme, um deren Widerstandsfähigkeit gegenüber KI-spezifischen Risiken, wie adversariellen Angriffen und Datensatzmanipulationen, zu validieren.

Netzwerksicherheit. Weiterhin muss die Sicherheit von Netzwerken jederzeit gewährleistet sein. Die Netzwerke, in denen IKT- und damit auch KI-Systeme betrieben werden, sind u. a. in Abhängigkeit von der Kritikalität zu segmentieren und gegen unbefugten Zugriff zu schützen. Es müssen spezielle Maßnahmen implementiert werden (z. B. Firewall-Regeln, Verschlüsselung der Netzwerkkommunikation), um die Integrität und Vertraulichkeit der Datenflüsse zu gewährleisten (Art. 9 DORA und Art. 13 RTS RMF).

Finanzunternehmen erstellen Richtlinien und Verfahren für die Netzwerksicherheit, in denen spezielle Anforderungen für KI-Systeme berücksichtigt werden sollten. Die Kritikalität und Bedeutung von KI-Systemen fließen im Idealfall ebenfalls angemessen in die Trennung und Segmentierung der IKT-Systeme und Netzwerke ein (Art 13 lit. a RTS RMF). Da KI-Systeme vertrauliche Informationen verarbeiten können, ist es empfehlenswert, den Zugang zu Netzwerken risikoorientiert physisch und/oder technisch zu schützen.

Zur Härtung des Netzwerkes haben sich einige Sicherheitsmaßnahmen etabliert:

- Einsatz von Web-Proxies zur Untersuchung, Überwachung und Absicherung des Netzwerkverkehrs unter Berücksichtigung sämtlicher unverschlüsselter und verschlüsselter Kommunikation.
- Einsatz von Web-Application-Firewalls und API-Gateways zur Filterung von Angriffen auf Anwendungsebene.
- Schutz vor DDoS-Angriffen aus dem Internet auf KI-Systeme.
- Einrichtung von Zugängen zu VPN und Desktop-Fernzugriff und automatische Terminierung von Fernsitzungen.

Berechtigungsmanagement. Strikte Zugriffskontrollen und Berechtigungsmanagementsysteme sind notwendig, um unbefugten Zugriff auf sensible Daten zu verhindern. Entsprechend sind Authentifizierungs- und Autorisierungsverfahren einzusetzen, um den Zugang zu sensiblen Daten zu beschränken. Die umfassende Protokollierung aller Datenzugriffe und -änderungen ist notwendig, um verdächtige Aktivitäten zu identifizieren und zu verfolgen. Rollenbasierte Zugriffssteuerungssysteme (RBAC) können sicherstellen, dass Benutzer nur auf die Daten zugreifen können, die sie für ihre Arbeit benötigen (Art. 9 Abs. 4 lit. c DORA und Art. 21 lit. a RTS RMF).

Überwachung und Aufzeichnungen. Es empfiehlt sich, alle relevanten Ereignisse und Aktivitäten in KI-Systemen aufzuzeichnen. Die Protokollierung dient der Nachvollziehbarkeit, der Erkennung anomaler Aktivitäten und wird typischerweise gegen Manipulation geschützt (Art. 12 RTS RMF).

Artikel 24, 25 DORA

Testen der digitalen operationalen Resilienz

Finanzunternehmen implementieren ein solides und umfassendes Programm für das Testen der digitalen operationalen Resilienz als integraler Bestandteil des IKT-Risikomanagementrahmens.

IKT-Änderungsmanagement. Für Notfalländerungen an KI-Systemen können spezielle Freigabe- und Bewertungsprozesse definiert werden, die ein kurzfristiges Einspielen von Änderungen ermöglichen (Art. 17 Abs. 1 lit. f und g RTS RMF).

Resilienztests. Zudem sieht DORA vor, dass Finanzunternehmen regelmäßige Resilienztests durchführen (Art. 24 und 25 DORA). Erweiterte Tests auf Basis von TLPT werden an dieser Stelle nicht betrachtet (Art. 26 bis 27 DORA).

Um die Cybersicherheit von KI-Systemen sicherzustellen, haben sich zahlreiche Maßnahmen bewährt. Diese schließen ein:

- Kontinuierliche Echtzeit-Überwachung von KI-Systemen auf anomales Verhalten (z. B. unerwartete Entscheidungsmuster).
- Einrichtung von Schutzmechanismen gegen adversarielle Eingaben, z. B. durch Filtermechanismen.
- Protokollierung des relevanten Outputs des KI-Systems und der API-Aufrufe für spätere Analysen.
- Regelmäßige Updates des KI-Systems, um neuen Bedrohungen zu begegnen.
- Notfallplan für den Umgang mit Sicherheitsvorfällen im KI-System.

2. Datensicherheit

Finanzunternehmen verarbeiten vertrauliche Daten. Es ist wichtig, dass sie diese Daten schützen, auch wenn sie KI-Systeme einsetzen.

Datenklassifikation, -übertragung und -verschlüsselung. Die wichtigste Grundlage für die sichere Verarbeitung und Speicherung von Unternehmensdaten ist die Klassifikation. Dabei sind die Anforderungen an die Vertraulichkeit, Integrität und Verfügbarkeit zu berücksichtigen. Diese Einstufung bestimmt, wie und wo Daten verarbeitet und gespeichert werden können (Art. 5 Abs. 2 lit. b RTS RMF). Eine derartige Klassifikation ist eine sinnvolle Ergänzung für den Betrieb von KI-Systemen in der Cloud. DORA sieht vor, dass Finanzunternehmen alle ihre Daten entsprechend klassifizieren (Art. 11 Abs. 2 lit. k RTS RMF).

KI-Systeme werden gemäß den festgelegten Vorgaben verschlüsselt, also nach entsprechender Datenklassifizierung und IKT-Risikobewertung. Dies umfasst Daten, die gespeichert, übermittelt oder, soweit dies erforderlich ist, benutzt werden. Sofern eine Verschlüsselung während der Verarbeitung nicht erfolgen kann, sind mindestens andere Maßnahmen, wie Verarbeitung der Daten in einer getrennten und geschützten Umgebung, notwendig (Art. 6 Abs. 2 RTS RMF). Der sichere Umgang mit kryptografischen Schlüsseln, etwa zur Absicherung von Kommunikationskanälen zwischen KI-Systemkomponenten, ist empfehlenswert. DORA fordert hier explizit den Einsatz geeigneter Verschlüsselungsverfahren, kryptografische Kontrollen und Schlüsselmanagement (Art. 6 und 7 RTS RMF).

Auch die Übertragung von Daten zwischen KI-Systemkomponenten, sowie zwischen KI-Systemen und externen Diensten, wird typischerweise gesichert, um z. B. vor Datenlecks zu schützen. Es sind Verfahren zur Überwachung und Bewertung der Einhaltung dieser Anforderungen einzurichten (Art. 14 RTS RMF).

Es haben sich Verhaltensweisen herausgebildet, die unternehmensindividuell angepasst werden können:

- Verschlüsselung und Signierung von Modellen, um unautorisierte Änderungen zu verhindern.
- Nutzung sicherer Umgebungen (z. B. Container).
- Implementierung eines Zero-Trust-Modells für den Zugriff auf KI-Dienste.
- Absicherung von KI-Systemen gegen Injection-Angriffe, Rate-Limiting zur Verhinderung von Denial-of-Service-Angriffen.

Datenqualität. Typischerweise umfasst der Begriff der Datenqualität mehrere Aspekte, wie etwa Vollständigkeit, Genauigkeit, Gültigkeit, Konsistenz, Angemessenheit bzw. Repräsentativität und Integrität.

Während die Sicherstellung der Datenintegrität ein Schutzziel von DORA ist, enthält die Verordnung keine Regelungen für andere Aspekte der Datenqualität. Eine hohe Datenqualität v. a. bei Trainingsdaten ist dennoch eine wesentliche Voraussetzung für den Einsatz von Künstlicher Intelligenz, da die Genauigkeit, Verlässlichkeit und Leistung von KI-Systemen davon abhängen. Vorgaben zur Datenqualität sind typische Bestandteile von Governance-Strukturen zum Einsatz von KI. Für Besonderheiten bei KI-Assistenten und LLMs siehe die Fallstudie.

3. Meldung schwerwiegender IKT-bezogener Vorfälle

Meldung schwerwiegender IKT-bezogener Vorfälle. Finanzunternehmen sind zur zeitnahen Meldung schwerwiegender IKT-bezogener Vorfälle an die zuständigen Behörden verpflichtet (Art. 19 DORA). Diese Meldeplicht kann auch Vorfälle in KI-Systemen umfassen. Schnelle Reaktionsmechanismen sind notwendig, da z. B. Manipulationen von KI-Systemen in kurzer Zeit erhebliche operative Schäden verursachen können.

Artikel 17 DORA

IKT-Vorfälle

Finanzunternehmen bestimmen einen Prozess für die Behandlung IKT-bezogener Vorfälle, richten diese ein und wenden sie an, um IKT-bezogene Vorfälle zu erkennen, zu behandeln und zu melden.

Unternehmen müssen eine Richtlinie für den Prozess zur Behandlung IKT-bezogener Vorfälle erstellen. Diese Richtlinie geht idealerweise auch auf KI-Systeme ein. Ebenso sollen technische, organisatorische und operative Mechanismen zur Unterstützung des Prozesses für die Behandlung IKT-bezogener Vorfälle implementiert und betrieben werden (Art. 22 RTS RMF).

Vorfälle, die durch oder im Zusammenhang mit IKT-Systemen, wie bspw. KI-Systemen, auftreten, die Sicherheit der Netzwerk- und Informationssysteme beeinträchtigen und die IT-Schutzziele von Daten verletzen oder negative Auswirkungen auf erbrachte Dienstleistungen haben, sind als IKT-bezogene Vorfälle zu erfassen (Art. 17 DORA). Hierbei empfiehlt es sich, Vorfälle im Zusammenhang mit KI-Systemen entsprechend zu kennzeichnen.

Bei der Nutzung von KI-Systemen in der Cloud empfiehlt es sich, Vereinbarungen zur Meldung von IKT-Vorfällen, welche in KI-Systemen entstehen, zu treffen. Außerdem sollten Unternehmen ausreichende und qualifizierte unternehmensinterne Ressourcen zur Bewertung und Reaktion bereitstellen.

Bewährte Maßnahmen für die Behandlung von Vorfällen schließen ein:

- Identifizierung von KI-spezifischen Bedrohungen (z. B. Manipulation von Trainingsdaten).
- Erkennen von IKT-Vorfällen, welche sich aus KI-Systemen ergeben, um z. B. bei Modellfehlern, Datenverlust oder Performance-Problemen unmittelbar handeln zu können.
- Auswirkungsanalyse (z. B. Datenverlust) von IKT-Vorfällen, welche aus KI-Systemen entstehen, sowie Einstufung der Schweregrade.
- Integration derartiger Vorfälle in das allgemeine Incident-Response-Konzept.

Nach solchen Vorfällen empfiehlt sich eine detaillierte Ursachenanalyse, um systematische Schwachstellen in KI-Modellen und -Prozessen zu identifizieren und zu beheben. Diese Ergebnisse können in kontinuierlichen Verbesserungsmaßnahmen genutzt werden, um wiederholte IKT-Vorfälle in KI-Systemen nachhaltig zu verhindern.

VI. Schlussbetrachtung und Ausblick

Die Nutzung von KI-Systemen bietet deutliches Potenzial für Finanzunternehmen, nicht zuletzt durch effizientere interne Prozesse und eine genauere Steuerung von Risiken. Jedoch steigen mit der intensiveren Nutzung von KI auch die damit verbundenen Risiken.

Daher ist es notwendig, dass Finanzunternehmen zunächst geeignete Governance- und Organisationsstrukturen definieren, um IKT-Risiken aus KI-Systemen zu mitigieren. DORA macht ausreichende Vorgaben, um einen wirksamen IKT-Risikomanagementrahmen für solche Systeme einzurichten.

Die Entwicklung und das Testen von KI-Systemen stellen Unternehmen vor besondere Herausforderungen. Dies liegt an der oftmals hohen Komplexität der mathematischen Modelle, an der Menge der zu verarbeitenden Daten sowie an der Anbindung von Software von IKT-Drittdienstleistern. Selbst- und fremdentwickelte KI-Systeme sind nach denselben Standards zu analysieren und zu testen.

Für den Betrieb von KI-Systemen sind entsprechende Prozesse zu definieren. Diese sollen Risiken und IKT-Vorfälle während des gesamten Lebenszyklus eines KI-Systems identifizieren und Gegenmaßnahmen einleiten. Da KI-Systeme häufig in Cloud-Umgebungen ausgeführt werden, sind die Vorgaben der Nutzung von IKT-Drittdienstleistern entsprechend zu interpretieren.

Da Finanzunternehmen KI-Systeme auf vertraulichen Daten einsetzen, typischerweise Kundendaten, sind die Systeme gegen unbefugte Zugriffe und Datenabflüsse zu schützen. Der Cyber- und Datensicherheit kommt eine hohe Bedeutung zu, die ggf. signifikante Anstrengungen der Unternehmen für die Klassifizierung der Daten in Vertraulichkeitsstufen verlangt. Für schwerwiegende IKT-bezogene Vorfälle sind geeignete Meldeprozesse einzurichten.

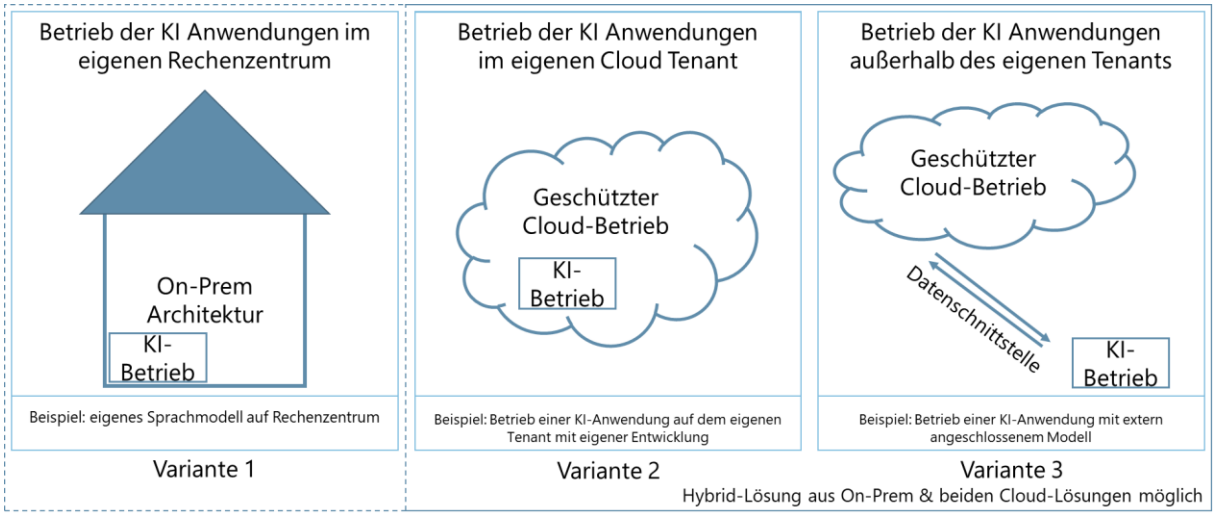
Besonders relevant sind diese Anforderungen für KI-Assistenten, die häufig in Standardsoftware enthalten sind. Die Implementierung von KI-Assistenz-Systemen in einem Finanzdienstleistungsunternehmen erfordert eine ganzheitliche Sicherheitsstrategie, die technische, organisatorische und regulatorische Maßnahmen umfasst.

Fallstudie—Betrieb eines LLM-basierten KI-Assistenten

Ein Unternehmen im Finanzdienstleistungssektor plant die Implementierung eines LLM-basierten KI-Assistenten zur Unterstützung der Mitarbeiter, um Texte, E-Mails und Präsentationen zu erstellen. Da das Unternehmen mit sensiblen Finanz- und Kundendaten arbeitet, müssen strenge Sicherheits- und Compliance-Anforderungen erfüllt werden.

Diese Fallstudie beschreibt typische Risiken bei der Implementierung einer solchen KI-Anwendung. Dabei werden drei Varianten für die Infrastruktur untersucht: die On-Premise-Lösung und zwei Varianten mit Cloud-Betrieb, nämlich die Nutzung einer KI-Anwendung im eigenen Tenant und die Nutzung einer KI-Anwendung außerhalb des eigenen Tenants. Abbildung 2 stellt die Varianten konzeptionell dar.

Abbildung 2: Archetypische Infrastrukturvarianten



Die Fallstudie zeigt eine beispielhafte Diskussion ausgewählter IKT-Risiken, die beim Einsatz von KI-Assistenten beobachtet werden können. Sie formuliert keine aufsichtlichen Erwartungen. Alle in der Fallstudie dargestellten Maßnahmen sind exemplarisch. Demgegenüber müssen Finanzunternehmen diejenigen Maßnahmen umsetzen, die für ihre spezifische Risikosituation am besten geeignet sind.

Die konkret in den Finanzunternehmen genutzten IKT-Assets und die Infrastruktur kann auch eine Kombination aus Elementen der drei Varianten darstellen. Die nachfolgende Diskussion der Risiken ist dann entsprechend anzupassen.

Unabhängig davon gibt es Risiken, die in jeder Infrastrukturvariante von Bedeutung sind. Entlang des KI-Lebenszyklus lauten wesentliche Beispiele wie folgt:

1. Datenbeschaffung und –aufbereitung: Ein wesentliches Risiko besteht darin, dass die KI-Anwendung Daten verarbeitet, die nicht für die Anwendung freigegeben wurden. Daher ist eine umfassende Klassifizierung der Daten nach Vertraulichkeitsstufen unerlässlich. Etwas ausführlicher können ausgewählte Maßnahmen wie folgt kategorisiert werden.

Datenklassifizierung und Zugriffskontrolle:

- Wenn möglich automatisierte Erkennung und Klassifikation vertraulicher Daten.
- Intensive Schulung zur Nutzung der KI-Anwendung.
- Umsetzung einer Zero-Trust-Policy, bei der die KI-Anwendung nur autorisierte Daten abrufen kann sowie Einführung eines rollenbasierten Modells für den Datenzugriff für Nutzer der KI-Anwendung.

Datenminimierung und Maskierung:

- Nutzung von differentieller Privatsphäre soweit möglich, um personenbezogene Daten vor möglichen Rückschlüssen im Rahmen von Mehrfachabfragen zu schützen (Genauigkeit von Antworten zu Anfragen an Datenbanken maximieren, unter Minimierung der Wahrscheinlichkeit, die zur Beantwortung verwendeten Datensätze identifiziert werden können).
- Tokenisierung von sensiblen Daten vor der Verarbeitung durch den KI-Assistenten.

Erkennung und Vermeidung von Bias in den Daten:

- Datenprüfung und –bereinigung vor dem Einsatz u. a. mittels automatischer Validierung oder manuellen Stichproben
- Schulung von Nutzern hinsichtlich der Datenauswahl und -aufbereitung

2. Modellentwicklung und Training: Beim (Nach)Trainieren von LLMs besteht das Risiko, dass manipulierte Daten eingesetzt werden, was zu unerwartetem bzw. veränderten Verhalten des Modells führen kann. Als Maßnahmen kommen hier in Betracht:

- Sicherstellen, dass der KI-Assistent nur mit geprüften und validierten Unternehmensdaten (nach)trainiert wird, um die sog. Datenvergiftung (Data Poisoning) zu vermeiden. Hilfreich kann dabei auch die Einrichtung eines Data Governance Teams sein, das kontrolliert, welche Daten für KI-Verarbeitung freigegeben werden.
- Sicherstellen, dass bei Retrieval Augmented Generation diejenigen Daten, die als Wissensbasis (sog. Kontext oder Grounding) dem LLM beigegeben werden, ebenfalls geprüft und validiert sind, um die sog. Wissensvergiftung (Knowledge Poisoning) zu vermeiden.
- Versionskontrollsysteme für Modelle, um ein Roll-Back bei z. B. Fehlfunktion oder aufgrund von fehlerhaftem Training zu ermöglichen.
- Für jede Art der Anwendung von generativer KI empfiehlt es sich, eigene Testverfahren zu entwickeln und zu nutzen, die die speziellen Anforderungen des Use Case reflektieren.

Analog zur Datenvergiftung gibt es die sog. Modellvergiftung. Dieses Risiko tritt besonders bei Open Source-Modellen auf oder Modellen, die über geteilte Repositiorien bezogen werden. Hier versuchen Angreifer, das Verhalten eines trainierten Modells oder gar die Struktur des Modells zu verändern. Best-Practice-Maßnahmen zur Risikomitigierung schließen ein:

- Überprüfung des Quellcodes von öffentlich zugänglichen Modellen auf Backdoors und Schadcode.
- Sicherstellen der Integrität von Repositorien und Modell-Anbietern.

Weiterhin können – auch bei Nutzung eines vortrainierten Modells – Angreifer versuchen, Backdoor-Angriffe oder adversariale Manipulationen umzusetzen. Potenzielle Maßnahmen dagegen sind etwa:

- Implementierung eines Auditing-Frameworks, das alle Anfragen an den KI-Assistenten protokolliert und überprüfbar macht. Hilfreich sind dabei Explainable AI-Techniken zur Erklärung von Ausgaben des KI-Assistenten.
- Einsatz von anomalieerkennenden Sicherheitsmechanismen, um ungewöhnliche Modellantworten oder verdächtige Abfragen zu identifizieren.
- Regelmäßige Penetrationstests für die KI-Assistenten-Umgebung, um Schwachstellen frühzeitig zu erkennen sowie Red-Teaming-Ansatz, bei dem Sicherheitsexperten versuchen, das Modell zu manipulieren und Schwachstellen zu finden.

3. Modellbereitstellung und –integration: Auch in dieser Phase des KI-Lebenszyklus entstehen IKT-Risiken. Vornehmlich, wenn der KI-Assistent nicht sicher implementiert wird, können Angreifer unautorisierte Zugriffe auf interne Systeme erhalten oder Modellinformationen exfiltrieren. Als geeignet haben sich dabei folgende Maßnahmen herausgestellt.

Deployment-Strategie:

- Der KI-Assistent wird über eine isolierte Cloud-Umgebung bereitgestellt, um direkten Internetzugriff zu verhindern.
- Nutzung von Containerisierung, um die KI-Umgebung von anderen IKT-Systemen zu trennen.

Zugriffskontrolle und Authentifizierung:

- Implementierung von Multi-Faktor-Authentifizierung (MFA) für KI-Assistent-Nutzer.
- Nutzung von Conditional Access Policies, sodass nur autorisierte Benutzer mit firmeneigenen Geräten den KI-Assistent nutzen können.
- Kritische Prüfung von Schnittstellen und Zugriffsregelungen zu weiteren Unternehmenssystemen unter Einbindung des Human-in-the-Loop Prinzips (z. B. Freigabe durch Eigentümer der Daten).

API-Sicherheit:

- Alle API-Abfragen zwischen Unternehmenssystemen und KI-Assistent laufen über verschlüsselte Schnittstellen.
- Rate-Limiting und DDoS-Schutz durch Cloud-Sicherheitsmechanismen, um Überlastungsangriffe zu verhindern.

4. Betrieb und Nutzung: Die Einbindung von LLMs in Unternehmensprozesse führt zu zahlreichen Risiken. Neben Herausforderungen, die nicht unmittelbar IKT-relevant sind (wie z. B. Halluzinationen, also die überzeugende Formulierung vermeintlich korrekter aber tatsächlich falscher Aussagen durch das Modell) sind verschiedene IKT-Risiken relevant. Darunter fallen folgende wesentliche Themen:

Erstens die Extraktion/Offenlegung von sensiblen Geschäftsinformationen aus nachtrainierten öffentlich zugänglichen LLMs. Hier können durch entsprechende Aufforderungen (sog. Prompts) ausgewählte und vertrauliche Informationen aus dem Trainingsdatensatz wiedergegeben werden. In diesem Zusammenhang besteht das Risiko, dass das LLM durch maliziöse Prompts (sog. Prompt Injection) zu nicht geplanten Verhaltensweisen gezwungen wird, z. B. durch einen Verweis auf Internetseiten, die verborgene Anweisungen an das LLM enthalten. Um diesen Risiken zu begegnen, bieten sich u. a. folgende Maßnahmen an:

- Nutzung von Erklärbarkeits-Werkzeugen, die es Nutzern ermöglichen, nachzuvollziehen, warum die KI eine bestimmte Antwort gibt.
- Bereitstellung von KI-Sicherheits-Trainings für Mitarbeiter, um Fehlinterpretationen und unachtsame Nutzung zu minimieren.
- Untersuchung, welche Prompts zu welchen Effekten führen und Setzen geeigneter Beschränkungen.

Zweitens kann ein kompromittierter oder fehlerhaft arbeitender KI-Assistent vertrauliche Informationen an die Nutzer geben, die eigentlich keinen Zugriff auf diese Daten haben. In diesem Kontext können Prozesse zum Monitoring der KI-Anwendung und zur Anomalieerkennung eingesetzt werden, z. B.

- Implementierung einer situationsbezogenen Überwachung der KI-Interaktionen.
- Identifizierung von anormalen Aktivitäten innerhalb der KI-Assistenz-Umgebung.

Drittens können API-Zugänge zu LLMs Angreifern die Möglichkeit bieten, auf Systeme zuzugreifen, die dem LLM angeschlossen sind und somit vertrauliche Informationen abzuschöpfen. Hier sollten Finanzunternehmen neben der geeigneten Zuordnung von Zugriffsrechten auch die Isolierung von LLMs betrachten. Weitere in Frage kommende Maßnahmen können sein:

- Eine potenzielle Risikomitigierung kann in Einschränkungen der LLM-Nutzung für KI-Systeme, die kritische oder wichtige Funktionen unterstützen, liegen.
- Einführung einer menschlichen Überprüfungspflicht (Human-in-the-Loop) für sicherheitskritische KI-Antworten/-Empfehlungen.

5. Wartung, Updates und Incident Response: Insbesondere bei der Nutzung von Open-Source-LLMs ist zu beachten, dass veraltete oder falsch konfigurierte Software-Versionen zu Sicherheitslücken führen können. Ausgewählte Maßnahmen, um diesem Risiko zu begegnen schließen die folgenden Themen ein.

Regelmäßige Sicherheitsupdates:

- Automatische Updates für den KI-Assistenten automatisiert gesteuert über ein geeignetes Werkzeug sowie Nutzung von Patch-Management-Tools, um Sicherheitslücken schnell zu schließen.
- Einführung eines Security Incident Response Plans (SIRP) für KI-Assistenten-spezifische Sicherheitsvorfälle.
- Simulation von Cyberangriffen, um die Reaktionszeiten des Unternehmens zu testen.

Vorgaben in der Governance:

- Regelmäßige externe Audits, um sicherzustellen, dass der KI-Assistent weiterhin mit Unternehmensrichtlinien übereinstimmt.
- Zentrale Erfassung Governance Aspekte im Rahmen eines Compliance Dashboards, das Transparenz über die Nutzung liefert (bspw. Modellversion, Risikobewertungen, Verantwortlichkeiten)
- Es erscheint dabei sinnvoll, neben allg. gültigen Anforderungen, auch technologische Besonderheiten für KI-Systeme wie etwa on-premise und/oder in der Cloud betriebene Systeme zu berücksichtigen.

6. End-of-Life Management: Die unkontrollierte Weiterverwendung oder unsichere Stilllegung von KI-Assistenten kann dazu führen, dass historische Daten und Modelle missbraucht oder unbeabsichtigt öffentlich verfügbar (geleakt) werden. Finanzunternehmen müssen daher – wie bei allen IKT-Assets – auch die Stilllegung von LLMs und zugehörigen Datenquellen planen. Potenzielle Best-Practice-Maßnahmen zur Risikomitigierung sind wie folgt.

Sichere Stilllegung von KI-Assistenten:

- Löschung aller verwendeten Daten und historischer KI-Interaktionen gemäß DSGVO-Richtlinien.
- Nutzung von kryptografischem Wiping, um alle gespeicherten Unternehmensdaten sicher zu entfernen.
- Sperrung des KI-Assistenten für abgelaufene Benutzerkonten oder ehemalige Mitarbeiter.

Im Folgenden werden nun die Besonderheiten der ausgewählten Infrastrukturmodelle untersucht. Nicht betrachtet werden dabei die Investitionskosten und laufenden Kosten. Diese werden in jeder Variante in größerem Umfang entstehen, da entweder der Aufbau und Betrieb einer eigenen Infrastruktur erforderlich sind oder aber die sichere Konfiguration des Modells zu einer erhöhten Cloud-Vergütung führen kann.

Variante 1: On-Premise-Lösung

In dieser Variante verfügt das Finanzunternehmen über eigene Speicher- und Verarbeitungskapazitäten und betreibt (teilweise oder vollständig) selbstentwickelte KI-Software auf der eigenen Hardware. Dieser Fall liegt vor, wenn das Unternehmen ein LLM selbst implementiert, pflegt, trainiert und betreibt. Somit liegen alle Datenflüsse innerhalb der unternehmenseigenen Infrastruktur. Das Finanzunternehmen hat hier die volle Kontrolle über die IKT-Assets der KI-Anwendung, trägt jedoch alle Risiken, die sich aus der Entwicklung sowie dem Betrieb und vor allem der Wartung der Hard- und Software ergeben: hier stechen sowohl das strategische Risiko des Geschäftsmodells hervor als auch das operationelle Risiko aus dem IKT-System.

Die speziellen Risiken dieses Ansatzes schlagen sich im KI-Lebenszyklus vor allem in der Modellentwicklung und dem Betrieb nieder. Zunächst erfordern die Implementierung und Wartung insbesondere von Open-Source-Modellen ausreichende Kenntnisse und Fähigkeiten, sodass die Verfügbarkeit von geeigneten Mitarbeitenden ein strategisches Risiko ist.

Speziell beim Betrieb von selbst implementierten Modellen und Open Source-Modellen besteht das Risiko der Vulnerabilität gegenüber Cyber-Angriffen. Wenn der KI-Assistent nicht sicher implementiert und fortlaufend aktualisiert wird, können Angreifer unautorisierte Zugriffe auf interne Systeme erhalten oder Modellinformationen exfiltrieren. Eine geeignete Schutzmaßnahme ist ein engmaschig kontrolliertes Zugriffsmanagement, da die Auswirkungen unkontrollierter Zugriffe bei KI-Nutzung ungleich größer sind als ohne KI.

Weiterhin kommt dem Kapazitätsmanagement in dieser Infrastrukturvariante eine gesteigerte Bedeutung zu. Eine beliebige Skalierung der KI-Anwendung ist – im Gegensatz zu den Cloud-Varianten – nicht möglich, ohne die Hardware zu erweitern. Daher ist die Infrastruktur auf die zu erwartende Rechenleistung zu konfektionieren, regelmäßig zu überprüfen und ggf. anzupassen.

Variante 2: Cloud mit KI-Anwendung in eigenem Tenant

Hier verfügt das Finanzunternehmen über einen eigenen Tenant in einer Cloud-Umgebung. Innerhalb dieses Tenant wird ein KI-Assistent (etwa eine selbst implementiertes Open Source-Modell) betrieben. Somit finden die Datenflüsse zwar in der Cloud statt, jedoch ausschließlich innerhalb des Firmen-Tenant.

Das wesentliche Risiko in dieser Variante ist strategischer Natur, da eine hohe Abhängigkeit vom Modellbetreiber besteht, sofern keine Open Source-Implementierung genutzt wird. Als mitigierende Maßnahme kann die Nutzung mehrerer Modelle unterschiedlicher Anbieter gelten, was allerdings zu einem erhöhten Betriebs- und Wartungsaufwand führt. Weiterhin müssen Mitarbeitende über geeignete Fähigkeiten und Kenntnisse verfügen, insbesondere beim Einsatz von Open Source-Software.

Weiterhin bestehen Herausforderungen für das Kapazitätsmanagement, da eine Skalierung der KI-Anwendung innerhalb des Tenant nicht immer möglich ist. Als Gegenmaßnahme kommt eine frühzeitige Kapazitätsplanung in Betracht.

Variante 3: Cloud mit KI-Anwendung außerhalb des eigenen Tenant

In dieser Variante verfügt das Finanzunternehmen über einen Tenant in einer Cloud. Jedoch liegt die KI-Anwendung außerhalb dieses Tenant, sodass die Datenflüsse über die Begrenzung des Tenant hinausgehen. Dieser Fall tritt typischerweise ein, wenn ein großes Sprachmodell des Cloud-Anbieters mittels API angesteuert wird oder als KI-Assistent aus Standardsoftware (ggf. ohne Kenntnis des Benutzers) aufgerufen wird.

Das wesentliche Risiko in dieser Variante ist, dass Daten den Tenant verlassen und zum Anbieter des Modells fließen. Dem soll durch vertragliche sowie technische Maßnahmen entgegengewirkt werden. Ausgewählte technische Maßnahmen dafür sind:

- Einschränkung der Funktionalität (u. a. Begrenzung des Uploads) für bestimmte Nutzergruppen.
- Betrieb eines Filters, der prüft ob Inputdaten vertraulich sind (Governance Shield).
- Vorschaltung der Bedingungen der KI-Anwendung vor jeder Nutzung.

Eine weitere denkbare mitigierende Maßnahme ist, dass die KI-Anwendung nur eine geringere Datenfreigabe erhält, also z. B. nur weniger sensible Daten verarbeitet werden können. Dabei ist zu beachten, dass Nutzer diese Einstufungen nicht eigenmächtig überschreiten können.

Die technische Kontrolle von IT-Betrieb, Informationssicherheit usw. soll analog zur Cloud Aufsichtsmitteilung erfolgen.

Impressum

Herausgeber
Bundesanstalt für Finanzdienstleistungsaufsicht (BaFin)
Abteilung Cyberrisiken und Technologie im Finanzsektor
Graurheindorfer Straße 108, 53117 Bonn
Marie-Curie-Straße 24 – 28, 60439 Frankfurt am Main
www.bafin.de

Abkürzungsverzeichnis

A

Abs.	Absatz
API	Application Programming Interface
Art.	Artikel

B

BaFin	Bundesanstalt für Finanzdienstleistungsaufsicht
BAIT	Bankaufsichtliche Anforderungen an die IT
bzw.	beziehungsweise

C

CRR	Capital Requirements Regulation

D

d. h.	das heißt
DAST	Dynamic Application Security Testing
DLP	Data Loss Prevention
DOR	Digitale operationelle Resilienz
DORA	Digital Operational Resilience Act, Verordnung (EU) 2022/2554 des EuropäischenParlamentes und des Rates vom 14. Dezember 2022 über die digitale operationale Resilienz im Finanzsektor
DDoS	Distributed Denial of Service
DoS	Denial of Service

E

EU	Europäische Union

G

GPU	Graphics Processing Unit
GenAI	generative KI

I

IDS	Intrusion Detection System
IDV	Individuelle Datenverarbeitung
IKT	Informations- und Kommunikationstechnologie
IPS	Intrusion Prevention System
i. V. m.	in Verbindung mit

K

KAIT	Kapitalverwaltungsaufsichtliche Anforderungen an die IT

Kap.	Kapitel
KI	Künstliche Intelligenz
KI-VO	KI-Verordnung, Verordnung (EU) 2024/1689 des europäischen Parlaments und des Rates vom 13. Juni 2024 zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz

L

lit.	Buchstabe
LLM	Large Language Model, großes Sprachmodell

M

MaRisk	Mindestanforderungen an das Risikomanagement, BaFin-Rundschreiben
MFA	Multi-Faktor-Authentifizierung
ML	Maschinelles Lernen

O

OECD	Organisation for Economic Co-operation and Development

R

RBAC	Role-Based Access Control, rollenbasierte Zugangskontrolle
RTS	Regulatory Technical Standard
RTS RMF	Delegierte Verordnung (EU) 2024/1774 der Kommission vom 13. März 2024 zur Ergänzung der Verordnung (EU) 2022/2554 des Europäischen Parlaments und des Rates durch technische Regulierungsstandards zur Festlegung der Tools, Methoden, Prozesse und Richtlinien für das IKT-Risikomanagement und des vereinfachten IKT-Risikomanagementrahmens

S

s.	siehe
SAST	Static Application Security Testing
SIRP	Security Incident Response Plan
SLA	Service Level Agreement
sog.	sogenannt

U

u. a.	unter anderem
UAbs	Unterabsatz

V

v. a.	vor allem
VAIT	Versicherungsaufsichtliche Anforderungen an die IT
vgl.	vergleiche
VPN	Virtual Private Network

X

xAIT	Sammelbegriff für BAIT, KAIT, VAIT und ZAIT
XAI	Explainable AI, erklärbare KI

Z

ZAIT	Zahlungsdiensteaufsichtliche Anforderungen an die IT von Zahlungs- und E-Geld-Instituten
z. B.	zum Beispiel