

DER STAND DER KÜNSTLICHEN INTELLIGENZ IN DER ÖFFENTLICHEN RECHNUNGSPRÜFUNG

ERKENNTNISSE AUS
AUSGEWÄHLTEN LÄNDERN UND
DER EUROPÄISCHEN UNION

OECD-PUBLIKATIONEN ZUR KÜNSTLICHEN INTELLIGENZ

Mai 2026 **Nr. 58**



Finanziert von der
Europäischen Union

OECD-Publikationen zur Künstlichen Intelligenz

Der Stand der künstlichen Intelligenz in der öffentlichen Rechnungsprüfung

Erkenntnisse aus ausgewählten Ländern und der
Europäischen Union

Das vorliegende Dokument wird unter der Verantwortung des Generalsekretärs der OECD veröffentlicht. Die darin zum Ausdruck gebrachten Meinungen und Argumente spiegeln nicht zwangsläufig die offizielle Einstellung der Mitgliedstaaten der OECD wider.

Dieses Dokument wurde mit finanzieller Unterstützung der Europäischen Union erstellt. Die darin zum Ausdruck gebrachten Ansichten können in keiner Weise als die offizielle Auffassung der Europäischen Union angesehen werden.

Dieses Dokument sowie die darin enthaltenen Daten und Karten berühren weder den völkerrechtlichen Status von Territorien noch die Souveränität über Territorien, den Verlauf internationaler Grenzen und Grenzlinien sowie den Namen von Territorien, Städten oder Gebieten.

© OECD 2026



Namensnennung 4.0 International (CC BY 4.0)

Dieser Text wird unter der Lizenz „Creative Commons Namensnennung 4.0 International“ verfügbar gemacht. Mit der Nutzung dieses Werks erklären Sie sich mit den Bedingungen dieser Lizenz (<https://creativecommons.org/licenses/by/4.0/>) einverstanden.

Namensnennung – Sie müssen das Werk nennen.

Übersetzungen – Sie müssen auf die Originalfassung verweisen, Veränderungen an der Originalfassung kennzeichnen und den folgenden Text hinzufügen: *Bei Diskrepanzen zwischen der Originalfassung und der Übersetzung gilt nur der Text der Originalfassung.*

Bearbeitungen – Sie müssen auf die Originalfassung verweisen und den folgenden Text hinzufügen: *Dies ist eine Bearbeitung einer Originalfassung der OECD. Die in dieser Bearbeitung zum Ausdruck gebrachten Meinungen und Argumente sollten nicht als die offizielle Auffassung der OECD oder ihrer Mitgliedsländer ausgewiesen werden.*

Material Dritter – die Lizenz gilt nicht für Material Dritter, das im Werk enthalten ist. Wenn Sie solches Material nutzen, sind Sie dafür verantwortlich, die Genehmigung der Rechteinhaber einzuholen, und haften für etwaige Ansprüche wegen Rechtsverletzung.

Sie dürfen das OECD-Logo, die visuelle Identität der OECD und das Titelbild nicht ohne ausdrückliche Genehmigung verwenden, und Sie dürfen nicht den Eindruck erwecken, dass die OECD Ihre Nutzung des Werks billigt.

Alle Streitigkeiten, die sich aus dieser Lizenz ergeben, werden durch ein Schiedsverfahren beigelegt, das der Schiedsordnung des Ständigen Schiedshofs (Permanent Court of Arbitration – PCA) in der Fassung von 2012 folgt. Der Sitz des Schiedsgerichts ist Paris (Frankreich). Die Zahl der Schiedsrichter*innen beträgt eins.

Übersicht

Diese Publikation untersucht, wie öffentliche Rechnungsprüfungsinstitutionen den Einsatz von künstlicher Intelligenz (KI) erkunden, um die Aufsicht zu stärken und Prüfungsprozesse zu verbessern. Basierend auf Konsultationen mit 15 Institutionen aus 14 Ländern und der Europäischen Union werden neue KI-Anwendungen in Bereichen wie Anomalieerkennung, Dokumentenverarbeitung, Wissensmanagement und vorausschauender Risikobewertung untersucht. Die Ergebnisse zeigen, dass sich die Einführung von KI im Bereich der öffentlichen Rechnungsprüfung zwar noch in einem frühen Stadium befindet, die Experimentierfreude jedoch zunimmt und viele Institutionen KI im Rahmen umfassenderer digitaler Transformationsprozesse integrieren. Dennoch besteht weiterhin eine Lücke zwischen Pilotprojekten und einer skalierbaren operativen Umsetzung. Zentrale Herausforderungen umfassen fragmentierte Datensysteme, begrenzte interne technische Fachkenntnisse und sich weiterentwickelnde Governance-Rahmen. Die Stärkung der Daten-Governance, der digitalen Infrastruktur und der internen Entwicklungskapazitäten wird für Prüfungsinstitutionen von entscheidender Bedeutung sein, um künstliche Intelligenz verantwortungsvoll auszuweiten und dabei Transparenz, Rechenschaftspflicht und das öffentliche Vertrauen zu wahren.

Danksagungen

Diese Publikation wurde von Maria Eugenia Heyaca und Andrea Pallotta unter Mitwirkung von Helene Wells, Sara Minguez, Rita Guelzim und Matteo Frison erstellt, unter der Leitung von Elsa Pilichowski, Direktorin für Public Governance (GOV) der OECD, und Nejla Saula, kommissarische Leiterin der Abteilung für Korruptionsbekämpfung, Integrität und offene Verwaltung (GOV/ACIOG) der OECD. Neben diesen Mitwirkenden haben weitere OECD-Kolleg*innen den Bericht geprüft und wertvolle Anmerkungen beigesteuert, darunter Jamie Berryhill, András Hlács, Delphine Moretti und Alastair Swarbrick.

Diese Publikation wurde im Rahmen einer Zusammenarbeit mit den neun österreichischen Landesrechnungshöfen (LRH) von Burgenland, Kärnten, Niederösterreich, Salzburg, Steiermark, Tirol, Oberösterreich, Wien und Vorarlberg erstellt. Ihr Interesse an der internationalen Anwendung von künstlicher Intelligenz in der Prüfung hat den Schwerpunkt dieser Arbeit maßgeblich geprägt und spiegelt ihr umfassendes Engagement für neue Ansätze der öffentlichen Aufsicht wider.

Die Autor*innen möchten den folgenden Institutionen für ihre wertvolle Teilnahme an den Konsultationen danken:

- *Controladoria-Geral da União (CGU)*, Brasilien
- *Contraloría General de la República (CGR)*, Costa Rica
- *Cour des comptes*, Frankreich
- *Bundesrechnungshof*, Deutschland
- *Corte dei conti*, Italien
- Board of Audit and Inspection (BAI), Korea
- Europäischer Rechnungshof (EuRH), Europäische Union
- *Algemene Rekenkamer*, Niederlande
- *Riksrevisjonen*, Norwegen
- Accountant General's Department (AGD), Singapur
- Auditor-General's Office (AGO), Singapur
- National Audit Office (NAO), Vereinigtes Königreich
- Und den drei weiteren Institutionen, die ihr Wissen und ihre Erfahrung eingebracht haben, jedoch anonym bleiben wollten.

Die Autor*innen danken ihnen für ihre Erkenntnisse und Beiträge, die maßgeblich zur Ausgestaltung der Ergebnisse und Empfehlungen dieses Berichts beigetragen haben.

Das OECD-Sekretariat bedankt sich für die Unterstützung von Thomas Giacoletto, Projektkoordinator bei der Europäischen Kommission. Dieses Projekt wurde von der Europäischen Union über das Instrument für technische Unterstützung (TSI) finanziert und von der OECD in Zusammenarbeit mit der Europäischen Kommission über die Reform and Investment Task Force (SG REFORM) umgesetzt.

Diese Publikation profitierte von Konsultationen mit der OECD-Arbeitsgruppe für Integrität im öffentlichen Sektor und Korruptionsbekämpfung sowie dem OECD-Ausschuss für öffentliche Verwaltung.

Inhaltsverzeichnis

Übersicht	3
Danksagungen	4
Abkürzungen und Akronyme	8
1 Kurzfassung	10
KI in der öffentlichen Rechnungsprüfung: Eine im Entstehen begriffene, aber sich rasch entwickelnde Landschaft	10
Infrastruktur und Kapazität: Überbrückung der Kluft zwischen Pilotprojekten und dem produktiven Einsatz	10
Daten-Governance: Die unverzichtbare Grundlage für den Einsatz von KI	10
Gemeinsame Herausforderungen und neue bewährte Verfahren	11
2 Überblick und aktueller Stand	12
3 Anwendungsfälle	16
Anomalieerkennung und Risikobewertungen	16
KI zur Klassifizierung in Rechnungsprüfung und Aufsicht	19
Prädiktive und präventive Rechnungsprüfungsmodelle	20
Suche, Abruf und Wissensmanagement	22
Intelligente Dokumentenverarbeitung (IDP) und Datenextraktion	25
Generative KI für das Verfassen, Zusammenfassen und Übersetzen	26
Visuelle und räumliche Datenanalyse	29
4 Infrastruktur und Kapazität	36
Hardware-Infrastruktur	36
Softwareumgebung und -stack	38
Dateninfrastruktur und Governance	39
Entwicklungskapazitäten	40
5 Wirkungsmessung	44
6 Herausforderungen und strategische Überlegungen	46
Ethik, Erklärbarkeit und Verzerrung	46
Organisationskultur, Kompetenzen und Unterstützung der Einführung	47
Datenqualität als strategischer Wert	48
Governance und Sicherheit	48

Management des Veränderungstempos und Reduzierung der Abhängigkeiten von Drittanbietern	49
Ausgewogenheit zwischen Innovation, langfristiger Stabilität und Nachhaltigkeit	50
7 Gute Praktiken und Erfolgsfaktoren	52
Transparenz und Nachvollziehbarkeit	52
Einbindung von Nutzer*innen und Interessengruppen: Förderung von Innovationen von unten nach oben	53
Unterstützung durch Management und Führungsebene	54
Institutionelle Governance und Aufsicht	55
Wirksame Schulungen und Wissensaustausch	56
Strategische Partnerschaften	57
8 Zukünftige Entwicklungen	60
9 Schlussfolgerungen	62
Glossar	64
Literaturverzeichnis	67

Abkürzungen und Akronyme

KI	Künstliche Intelligenz
ML	Maschinelles Lernen
RAG	Retrieval-augmented Generation
CNN	Convolutional Neural Network
ETL	Extract-transform-load
GPU	Grafikprozessor
IDP	Intelligente Dokumentenverarbeitung
IoT	Internet der Dinge
IoU	Intersection over Union
NLP	Verarbeitung natürlicher Sprache
OCR	Optische Zeichenerkennung
RPA	Robotic Process Automation

1 Kurzfassung

KI in der öffentlichen Rechnungsprüfung: Eine im Entstehen begriffene, aber sich rasch entwickelnde Landschaft

Künstliche Intelligenz verändert die Aufsicht im öffentlichen Sektor grundlegend. Zwischen März und Juli 2025 befragte die OECD 15 Prüfungsinstitutionen in 14 Ländern sowie der EU, um den aktuellen Stand der KI-Einführung in der öffentlichen Rechnungsprüfung zu erheben. Die Ergebnisse deuten auf ein wachsendes institutionelles Engagement hin: Zwei Drittel verfügen über eine formelle KI-Strategie, 80% haben interne KI-Leitlinien, 87% bieten Schulungen für Mitarbeiter*innen an und ebenso viele haben mindestens ein Tool im aktiven Betrieb. Dennoch sind die Reifegrade sehr unterschiedlich, und viele Implementierungen befinden sich weiterhin im Pilotstadium.

Die Erkenntnisse deuten darauf hin, dass sich KI im Bereich der Prüfung zwar noch in der Entwicklung befindet, die Institutionen jedoch gezielt in die Voraussetzungen für eine breitere Nutzung investieren: Sie schaffen die Governance-Strukturen, Kompetenzen und technischen Grundlagen, die für eine umfassendere Einführung erforderlich sind.

Infrastruktur und Kapazität: Überbrückung der Kluft zwischen Pilotprojekten und dem produktiven Einsatz

KI-Anwendungen in der öffentlichen Rechnungsprüfung, die Bereiche wie Anomalieerkennung, prädiktive Modellierung, intelligente Dokumentenverarbeitung und generative KI abdecken, erfordern robuste, sichere und interoperable technologische Umgebungen. Die Konsultationen zeigen eine erhebliche Diskrepanz zwischen institutionellem Anspruch und verfügbarer Infrastruktur auf, wobei sich Hardware-, Software- und Datenökosysteme bei den teilnehmenden Institutionen deutlich unterscheiden.

Obwohl mehrere oberste Rechnungshöfen (supreme audit institutions, SAIs), bedeutende Fortschritte beim Aufbau KI-tauglicher Infrastrukturen erreicht haben, besteht weiterhin eine deutliche Kluft zwischen ersten Experimenten und den skalierbaren, produktionsreifen Umgebungen, die erforderlich sind, damit KI ihr Potenzial voll ausschöpfen kann. Die Überbrückung dieser Kluft wird in den kommenden Jahren zu den entscheidenden Investitionsprioritäten gehören, insbesondere da Prüfungsinstitutionen von Pilotprojekten zur Integration dieser Tools in zentrale Prozesse übergehen.

Daten-Governance: Die unverzichtbare Grundlage für den Einsatz von KI

Bei allen befragten Institutionen erwiesen sich Datenqualität, Zugänglichkeit und Governance als entscheidende Faktoren für den Erfolg von KI-Initiativen. Fragmentierte Systeme, mangelnde Standardisierung und geringe Interoperabilität schränken die Skalierbarkeit ein und beeinträchtigen das Vertrauen in die Ergebnisse von KI-Anwendungen.

Globale Standards verlangen zunehmend von öffentlichen Institutionen, Datenumgebungen bereitzustellen, die sicher, interoperabel und prüfbar sind. Da sich KI weiterentwickelt, werden Daten-Governance und Infrastruktur weiterhin die wichtigsten Treiber digitaler Innovation bleiben. Aufsichtsbehörden, die auf langfristige Interoperabilität und Governance statt nur auf kurzfristige Anwendungsfälle setzen, werden in den kommenden Jahren am ehesten in der Lage sein, anspruchsvollere Lösungen zu übernehmen. Da Rechnungsprüfungsbehörden zunehmend auf KI setzen, um Erkenntnisse und Kontrollen zu steuern, wird die Integrität der Daten, auf denen diese Tools basieren, noch entscheidender werden.

Gemeinsame Herausforderungen und neue bewährte Verfahren

Die teilnehmenden Institutionen weisen eine große Bandbreite an digitaler Reife, Ressourcen und Organisationsmodellen auf. Trotz dieser Vielfalt treten gemeinsame Hindernisse immer wieder auf: fragmentierte Datensysteme, begrenzte technische Fachkenntnisse im eigenen Haus und das Fehlen gemeinsamer Kennzahlen zur Messung der Auswirkungen von KI. Der Bericht dokumentiert dennoch echte Fortschritte, grenzüberschreitende Zusammenarbeit, die Einbindung des gesamten Personals sowie die Entstehung interner Governance-Rahmen, die auf einen verantwortungsvollen Einsatz von KI zugeschnitten sind.

Trotz dieser Einschränkungen gibt es deutliche Anzeichen für Innovation und institutionelles Engagement in den teilnehmenden Organisationen. Für die Zukunft wird es entscheidend sein, weiterhin in die interne Entwicklungskapazität, in robuste Rahmen für die Daten-Governance sowie in den internationalen Wissensaustausch zu investieren, um sicherzustellen, dass KI-Anwendungen in der öffentlichen Rechnungsprüfung einen wirkungsvollen Beitrag leisten und gleichzeitig mit den Werten des öffentlichen Sektors – Transparenz, Rechenschaftspflicht und Vertrauen – im Einklang bleiben.

2 Überblick und aktueller Stand

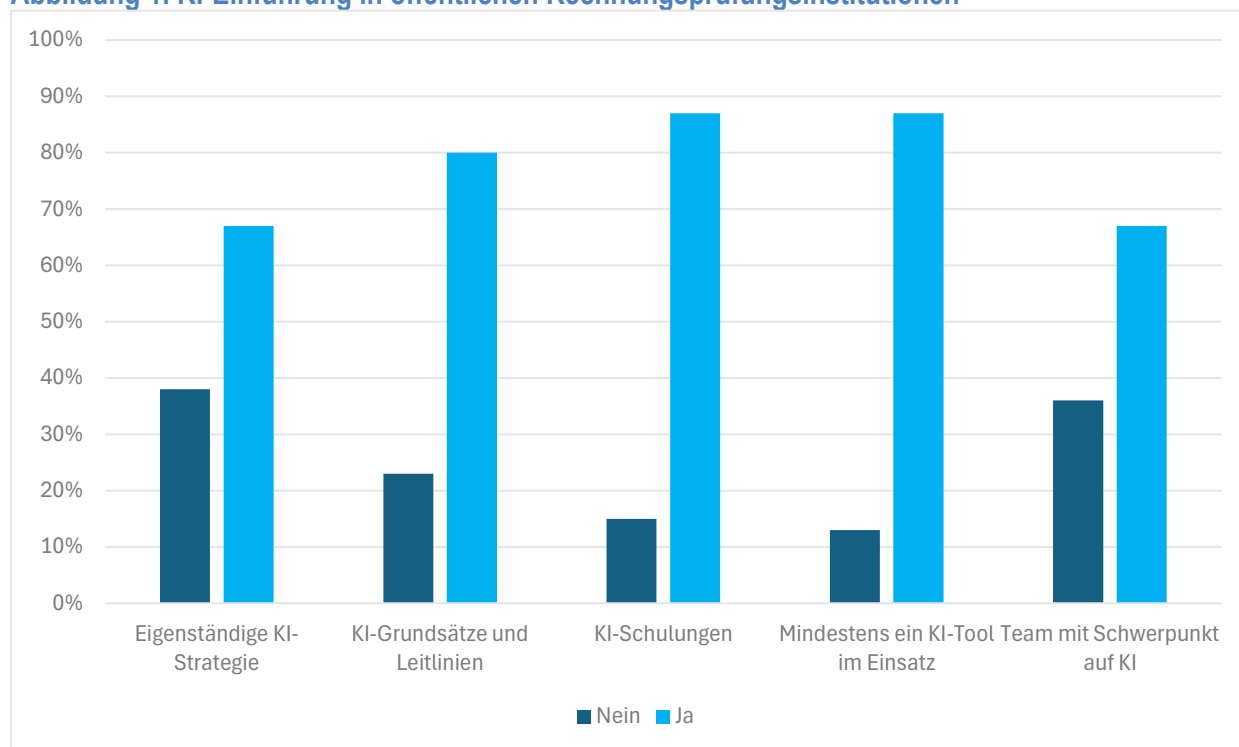
Künstliche Intelligenz (KI) verändert den öffentlichen Sektor rasant und bietet neue Tools für Aufsicht, Effizienz und Rechenschaftspflicht. Tatsächlich hat die Arbeit der OECD (2025^[1]) ergeben, dass verbesserte Rechenschaftspflicht und Anomalieerkennung zu den wichtigsten Vorteilen des Einsatzes von KI in der öffentlichen Verwaltung zählen. Öffentliche Rechnungshöfe untersuchen gemeinsam mit anderen öffentlichen Aufsichtsbehörden und relevanten Akteuren aus dem Privatsektor zunehmend, wie KI ihre Fähigkeit zur Erfüllung ihres Auftrags stärken kann. Obwohl das Potenzial erheblich ist, sind auch die Herausforderungen groß und reichen von Kapazitätsengpässen über ethische Bedenken bis hin zum Bedarf an klare Governance-Rahmen.

Zwischen März und Juli 2025 führte die Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD) Konsultationen mit 15 Institutionen in 14 Ländern (Brasilien, Costa Rica, Frankreich, Deutschland, Italien, Korea, den Niederlanden, Norwegen, Singapur, dem Vereinigten Königreich sowie drei weiteren Ländern, deren Institutionen anonym bleiben wollten) und der EU durch, um den aktuellen Stand der Einführung von KI in der öffentlichen Rechnungsprüfung zu bewerten. Die Gruppe bestand in erster Linie aus obersten Rechnungshöfen, umfasste jedoch auch eine multinationale Rechnungsprüfungsgesellschaft sowie eine öffentliche Aufsichtsbehörde. Die hier dargestellten Ergebnisse stützen sich zudem auf laufende Arbeiten und Partnerschaften der OECD mit anderen Institutionen und bieten damit eine umfassendere Perspektive auf neue Entwicklungen und Praktiken.

Dieser Bericht stellt das Hauptergebnis dieser Konsultationen und der damit verbundenen analytischen Arbeiten dar. Er fasst die wichtigsten Erkenntnisse und gewonnenen Einsichten zusammen und ist wie folgt gegliedert: Kapitel 2 untersucht eine Vielzahl von KI-Anwendungsfällen, die derzeit in verschiedenen Institutionen erprobt oder umgesetzt werden, und veranschaulicht die praktischen Anwendungen und Vorteile von KI im Prüfungssektor. Kapitel 3 untersucht die IT-Infrastruktur und den Kapazitätsaufbau, die diesen Bemühungen zur Einführung von KI zugrunde liegen. Kapitel 4 konzentriert sich auf Versuche, die Auswirkungen von KI-Initiativen zu messen, während Kapitel 5 die wichtigsten Herausforderungen und strategischen Überlegungen untersucht, denen sich Institutionen bei der Integration von KI-Technologien gegenüber sehen. Kapitel 6 hebt bewährte Verfahren und entscheidende Erfolgsfaktoren hervor, die sich aus den Erfahrungen der frühen Anwender*innen ergeben, und schließlich bietet Kapitel 7 eine Perspektive auf die Zukunft der KI im Bereich der öffentlichen Rechnungsprüfung, skizziert aufkommende Trends und potenzielle zukünftige Entwicklungen und schließt mit den Ergebnissen und Überlegungen des Berichts ab.

Die Ergebnisse zeigen, dass sich der Berufsstand der Rechnungsprüfung im Wandel befindet, wobei Experimentierfreude häufig mit strategischer Zielsetzung einhergeht, auch wenn Reifegrad und Integrationsniveau zwischen den Institutionen noch erheblich variieren. Abbildung 1 fasst die Ergebnisse der Umfrage zusammen, die die OECD zum aktuellen Stand der Einführung von KI in den befragten Organisationen durchgeführt hat.

Abbildung 1. KI-Einführung in öffentlichen Rechnungsprüfungsinstitutionen



Quelle: Konsultationen, die von der OECD zwischen März und Juli 2025 mit 15 Institutionen in 14 Ländern sowie der Europäischen Union durchgeführt wurden.

Zwei Drittel der befragten Institutionen (67%) geben an, eine formelle KI-Strategie eingeführt zu haben. Für Institutionen ohne eigenständige KI-Strategie übernehmen häufig Führungsvorgaben oder operative Ziele, die in Daten- oder IT-Abteilungen festgelegt werden, eine vergleichbare Funktion. In vielen Fällen sind KI-Initiativen auch in umfassendere Strategien zur digitalen Transformation oder Innovation eingebettet, was einen ganzheitlichen Blick auf den technologischen Wandel widerspiegelt.

Von den befragten Institutionen haben 80% interne KI-Leitlinien oder -Grundsätze eingeführt. Allerdings unterscheiden sich Art und Detaillierungsgrad dieser Dokumente erheblich. Einige Institutionen geben lediglich allgemeine oder zusammenfassende Erklärungen ab, in denen sie Grundsätze für den verantwortungsvollen Einsatz generativer KI skizzieren, während andere umfassendere Materialien entwickelt haben. Diese ausführlichen Dokumente können spezifische Verfahren, Anwendungsbeispiele oder maßgeschneiderte Empfehlungen enthalten, die auf die Bedürfnisse von technischem und nicht-technischem Personal zugeschnitten sind.

Der Aufbau von Kapazitäten erwies sich als zentrale Priorität. Beeindruckende 87% der Institutionen bieten derzeit KI-bezogene Schulungen für das allgemeine Personal an. Allerdings unterscheiden sich auch Umfang und Aufbau der Schulungen: In einigen Fällen beschränken sich die Lernmöglichkeiten auf Ad-hoc-Sitzungen, Seminare oder Konferenzformate, während andere Institutionen strukturiertere Programme eingeführt haben, die den Einsatz verschiedener KI-Techniken, einschließlich großer Sprachmodelle (Large Language Models, LLMs), auf unterschiedlich komplexen technischen Niveaus behandeln.

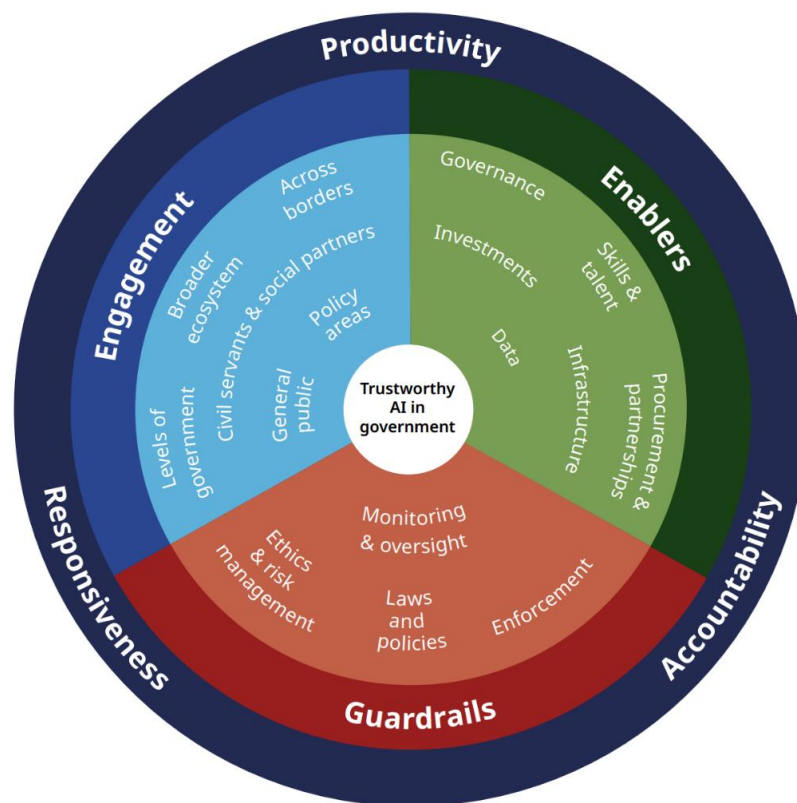
Der Einsatz von KI-Tools ist ebenfalls bereits weit fortgeschritten. 87% der Institutionen gaben an, derzeit mindestens ein KI-Tool produktiv im Einsatz zu haben, also in einen Arbeitsablauf integriert zu haben. Während dies für einige lediglich einen einzelnen Anwendungsfall bedeuten kann, etwa den Einsatz von KI zur Textzusammenfassung für die Unterstützung technischer Teams, haben andere bereits eine Reihe

von KI-gestützten Lösungen implementiert – ein Thema, das im nachfolgenden Abschnitt zu Anwendungsfällen näher erläutert wird. Dennoch befinden sich viele der eingesetzten Tools weiterhin in der Pilotphase, was einen allgemeinen Trend in den verschiedenen staatlichen Funktionen widerspiegelt (OECD, 2025^[1]). Dies deutet auf einen risikobewussten Ansatz hin, bringt jedoch auch verschiedene Herausforderungen mit sich, darunter den Aufbau einer geeigneten Infrastruktur, den Erwerb der erforderlichen Kompetenzen, die Messung der Auswirkungen zur Rechtfertigung weiterer Investitionen sowie die Entwicklung von Methoden, um eine vertrauenswürdige Nutzung von KI-Tools sicherzustellen.

Schließlich haben 67% der Institutionen eigene KI-Teams eingerichtet. Diese Teams spielen eine zentrale Rolle bei der Koordination der Entwicklung, Implementierung und Governance von KI-Systemen und fungieren häufig als interne Kompetenzzentren. Diese Teams bleiben überwiegend klein und bestehen in der Regel aus höchstens drei Personen.

Zusammen zeigen diese Ergebnisse ein wachsendes institutionelles Engagement für vertrauenswürdige KI, wenn auch mit unterschiedlich ausgeprägter strategischer Ausrichtung, Kapazität und praktischer Umsetzung. Wie die folgenden Abschnitte näher erläutern werden, schreitet der Einsatz von KI in öffentlichen Rechnungsprüfungsinstitutionen nicht nur voran, sondern diversifiziert sich auch und bietet wertvolle Erkenntnisse sowohl für aktuelle Anwender*innen als auch für Institutionen, die sich noch am Anfang ihrer KI-Entwicklung befinden. Diese Erkenntnisse spiegeln die im OECD-Bericht *Governing with Artificial Intelligence* (2025^[1]) dargelegten Beobachtungen wider und vertiefen sie, insbesondere im Hinblick auf die vorsichtige und fragmentierte Einführung von KI durch öffentliche Institutionen. Im Einklang mit den bisherigen Erkenntnissen der OECD bleiben Infrastrukturdefizite, regulatorische Unklarheiten und Fachkräftemangel die wichtigsten Hemmnisse. Der Schwerpunkt auf dem Kapazitätsaufbau, insbesondere durch gezielte Mitarbeiterschulungen, entspricht dem Aufruf der OECD, zentrale Voraussetzungen wie Personalentwicklung und digitale Kompetenzen zu stärken. Insgesamt sollen die Ergebnisse dieses Berichts den Prüfungsinstitutionen dabei helfen, wichtige Aspekte des OECD-Rahmens für vertrauenswürdige KI in der Verwaltung (Abbildung 2) besser zu verstehen und umzusetzen – und zwar aus einer Perspektive, die ihre spezifischen Rahmenbedingungen und Bedürfnisse besonders berücksichtigt.

Abbildung 2. OECD-Rahmenwerk für vertrauenswürdige Künstliche Intelligenz in der öffentlichen Verwaltung



Quelle: OECD (2025), *Governing with Artificial Intelligence*, https://www.oecd.org/en/publications/governing-with-artificial-intelligence_795de142-en.html.

3 Anwendungsfälle

Aufbauend auf den während der Konsultationen gewonnenen Erkenntnissen hebt dieser Abschnitt spezifische KI-Anwendungsfälle hervor, die von öffentlichen Rechnungsprüfungsinstitutionen entwickelt oder eingesetzt wurden. Diese praxisnahen Beispiele veranschaulichen, wie KI-Tools in verschiedenen Rechnungsprüfungsbereichen eingesetzt werden, von der Anomalieerkennung und Betrugsprävention bis hin zur Risikobewertung und Automatisierung von Arbeitsabläufen. Durch die Darstellung dieser Anwendungsbeispiele soll ein besseres Verständnis für die konkreten Vorteile geschaffen werden, die mit der Integration von KI in die Praxis der öffentlichen Rechnungsprüfung verbunden sind.

Zu den wichtigsten Anwendungsbereichen von KI gehören:

- Anomalieerkennung und Risikobewertungen
- KI zur Fallklassifizierung in Rechnungsprüfung und Aufsicht
- Prädiktive und präventive Rechnungsprüfungsmodelle
- Suche, Abruf und Wissensmanagement
- Intelligente Dokumentenverarbeitung (Intelligent Document Processing, IDP) und Datenextraktion
- Generative KI für das Verfassen, Zusammenfassen und Übersetzen
- Visuelle und räumliche Datenanalyse
- Weitere Anwendungen: interne Produktivitäts-, Kommunikations- und KI-Governance-Tools

Anomalieerkennung und Risikobewertungen

Rechnungsprüfungsinstitutionen nutzen zunehmend KI, um Anomalien zu erkennen, Risikostufen zu bewerten und den Einsatz von Prüfungsressourcen gezielter zu steuern. Neue Tools werden in Finanz-, Compliance- und Leistungsprüfungen integriert und können entweder im Rahmen einer kontinuierlichen Überwachung oder im Kontext spezifischer Prüfungsaufträge eingesetzt werden. In vielen Fällen wird KI eingesetzt, um traditionelle Stichprobenverfahren in der Prüfung zu ergänzen oder zu ersetzen, wodurch Transaktionen oder Einheiten, die einer weiteren Überprüfung bedürfen, präziser und in größerem Umfang identifiziert werden können.

Solche Anwendungen beinhalten typischerweise den Einsatz von Maschinelles Lernen (ML)-Algorithmen – wie Random Forest, Clustering-Techniken oder Autoencoder – zur Identifikation von Ausreißern in großen Datensätzen, einschließlich Finanz-, Beschaffungs- oder anderen Verwaltungsdaten. Sowohl überwachtes¹ als auch unüberwachtes Lernen[†] werden eingesetzt, um die Algorithmen zu trainieren, wobei letzteres tendenziell bevorzugt wird. Tatsächlich erfordert unüberwachtes Lernen keine gekennzeichneten Datensätze[†] und eignet sich besser dazu, bislang unbekannte oder sich entwickelnde Muster zu erfassen, die von überwachten Modellen, die auf historischen gekennzeichneten Daten trainiert

¹ Begriffe, die im Glossar definiert sind, werden bei ihrem ersten Auftreten im Text mit einem †-Symbol gekennzeichnet. Das Glossar enthält Definitionen für jeden gekennzeichneten Begriff.

wurden, nicht als Anomalien erkannt würden. Ein Beispiel für eine solche Anwendung findet sich beim Auditor-General's Office (AGO) von Singapur, wie in Kasten 1 dargestellt.

Ein typisches Beispiel für diesen Anwendungsfall ist das öffentliche Beschaffungswesen, das aufgrund der beträchtlichen finanziellen Interessen sowie der häufig engen Zusammenarbeit zwischen Beamten und privaten Anbietern von Natur aus anfällig für Bestechung, Absprachen und Preisabsprachen ist. Daher ist eine strenge Aufsicht unerlässlich, und der Einsatz von KI bei Beschaffungsprüfungen bietet erhebliches Potenzial, die Erkennung zu verbessern, die Effizienz zu steigern und die Integrität zu stärken, indem Risiken und Anomalien systematisch und in großem Umfang identifiziert werden. Die öffentliche Beschaffung ist ein Bereich, in dem der Einsatz von KI in der Prüfung zunehmend zunimmt, was auch auf die Verfügbarkeit großer Mengen digitaler Daten und die weit verbreitete Einführung von E-Beschaffungsplattformen zurückzuführen ist (wobei jedoch die Datenqualität, insbesondere hinsichtlich Vollständigkeit und Standardisierung, weiterhin eine Herausforderung darstellt). Beispielsweise hat der Rechnungshof der Republik Costa Rica (CGR) ML-gestützte Anomalieerkennung eingesetzt, um die Beschaffungsaufsicht zu stärken. Algorithmen wie Random Forest, Robust Components und Local Outlier Factor werden auf Daten aus dem nationalen E-Beschaffungssystem angewendet, um ungewöhnliche Beschaffungsmuster zu identifizieren. Dieser Ansatz verbessert traditionelle Stichprobenverfahren und liefert gezieltere Erkenntnisse über potenziell unregelmäßige Aktivitäten.

Bei Finanzprüfungen konzentrieren sich Anomalieerkennungsmodelle auf Hauptbuch- und Buchungseinträge, um potenziell betrügerische oder falsch klassifizierte Transaktionen zu identifizieren. Mehrere Institutionen haben KI-Systeme entwickelt, die KI-Risikoscores mit traditionellen Risikokennzahlen kombinieren und es den Prüfer*innen ermöglichen, risikobehaftete Transaktionen zu priorisieren, während das fachliche Urteil der Prüfer*innen im Entscheidungsprozess gewahrt bleibt. Die Institutionen haben diese Modelle in Prüfungsabläufe integriert und dabei aktive Feedbackmechanismen sowie Leistungskennzahlen wie Modellpräzision und Genauigkeit implementiert. Die Institutionen integrieren diese Tools zudem in Business-Intelligence-Plattformen, um – sofern möglich – eine Echtzeit-Risikobewertung zu unterstützen. Ein weiterer Schritt könnte darin bestehen, Feedbackschleifen mit den Dateneigentümern – insbesondere den Finanzministerien – weiterzuentwickeln, um die kontinuierliche Verbesserung der Systeme von zentralen staatlichen Stellen zu unterstützen.

SAIs untersuchen zudem fortschrittliche Verfahren wie graphenbasierte Modelle, um komplexe Anomalien zu identifizieren, die aus ungewöhnlichen Beziehungen oder Interaktionsmustern über mehrere Konten oder Einheiten hinweg entstehen. Durch die Darstellung von Buchhaltungsdaten als Netzwerke miteinander verbundener Elemente können diese Modelle unregelmäßige Strukturen aufdecken, die mit herkömmlichen Analysemethoden nur schwer zu erkennen sind. Neben graphenbasierten Ansätzen werden in der Praxis auch weitere fortschrittliche Methoden eingesetzt. Der Norwegische Rechnungshof *Riksrevisjonen* setzt beispielsweise neuronale Netzwerke ein, um die Arbeit der Prüfer*innen im Zusammenhang mit Betrugsrisiken zu unterstützen. Sie haben dieses Tool unter Berücksichtigung internationaler Prüfungsstandards entwickelt, sodass die Modelle nicht nur die Erkennung von Unregelmäßigkeiten wie doppelten Buchungen und unerklärlichen Stornierungen verbessern, sondern auch interpretierbare und visuell zugängliche Ergebnisse liefern, denen die Prüfer*innen vertrauen und die sie wirksam in ihre Prüfungsprozesse integrieren können (Huang et al., 2024^[2]).

Im Rahmen von Abschlussprüfungen besteht eine zentrale Herausforderung darin, echte Fehlangaben oder betrügerische Transaktionen von harmlosen Anomalien innerhalb der Buchungssätze zu unterscheiden, da viele Ausreißer durchaus ordnungsgemäße Geschäftsvorfälle sein können (zum Beispiel könnte eine ungewöhnlich große Buchung zum Jahresabschluss verdächtig erscheinen, jedoch lediglich eine legitime einmalige Anpassung nach einem Anlagenverkauf widerspiegeln). Ein Tool, das eine übermäßige Anzahl an falsch-positiven Ergebnissen erzeugt, verursacht Effizienzverluste, da Prüfungsressourcen für die Untersuchung legitimer Transaktionen aufgewendet werden. Umgekehrt beeinträchtigt eine hohe Rate an falsch-negativen Ergebnissen die Wirksamkeit der Prüfung, da potenziell

betrügerische Aktivitäten unentdeckt bleiben können. Dementsprechend sollten Modelle so kalibriert werden, dass das Gleichgewicht zwischen Effizienz bei Ermittlungen und Qualität der Prüfungsabsicherung optimal ist.

Die Verfügbarkeit umfangreicher, gekennzeichneteter Transaktionsdatensätze ist eine grundlegende Voraussetzung für das Training überwachter ML-Modelle zur Risikoklassifizierung. Angesichts der begrenzten Verfügbarkeit von gekennzeichneteten Daten setzen jedoch einige Institutionen unüberwachte Lernverfahren wie das K-Means-Clustering ein. Unüberwachte Lernmodelle gruppieren ähnliche Transaktionen und markieren solche, die vom typischen Verhalten abweichen, sodass SAs potenziell betrügerische Aktivitäten erkennen können, selbst wenn zuvor keine gekennzeichneten Betrugsfälle vorliegen.

Im Vereinigten Königreich hat das National Audit Office (NAO) ein Tool zur Anomalieerkennung entwickelt, das im Rahmen von Abschlussprüfungen eingesetzt wird, um risikobehaftete Transaktionen zu identifizieren. Dieses System verwendet eine Kombination aus analytischen Indikatoren und unüberwachten Methoden des maschinellen Lernens, um allen Geschäftsbüchern eine Risikobewertung zuzuweisen. Diese Geschäftsbücher, denen die höchsten Risikobewertungen zugewiesen werden, werden anschließend von Prüfer*innen überprüft und getestet. In Fällen, in denen Prüfer*innen feststellen, dass bestimmte risikobehaftete Geschäftsbücher nicht getestet werden müssen, sind sie verpflichtet, eine klare Begründung abzugeben oder eine unterstützende Analyse durchzuführen, um ihre Entscheidung zu untermauern.

Bei allen Implementierungen bestehen Herausforderungen darin, eine hohe Genauigkeit zu erreichen, die Datenqualität sicherzustellen, angemessene Stichprobengrößen festzulegen, KI-generierte Ergebnisse mit menschlichem Urteilsvermögen in Einklang zu bringen und das Modell-Driften[†] im Zeitverlauf zu überwachen. Dennoch stellt die Integration von Tools zur Anomalieerkennung und Risikobewertung einen bedeutenden Fortschritt in der Methodik der öffentlichen Rechnungsprüfung dar und erhöht sowohl die Effizienz als auch die Wirksamkeit.

Kasten 1. Auditor-General's Office (AGO) von Singapur: Einsatz von unüberwachtem maschinellem Lernen bei der Prüfung staatlicher Jahresabschlüsse

Die herkömmliche Anomalieerkennung in staatlichen Jahresabschlüssen basiert auf vordefinierten Regeln, die ungewöhnliche Transaktionen übersehen können, wenn diese nur geringfügig von den erwarteten Normen abweichen. Um diesen Prozess zu verbessern, setzte das AGO eine unüberwachte ML-Technik – „Isolation Forest“ – ein, um Anomalien in Buchungs- und Zahlungsdaten effektiver zu erkennen.

Bei diesem Verfahren werden die Buchhaltungsdaten in mehrere gezielte Gruppen unterteilt, sodass jede Teilmenge separat analysiert werden kann. Innerhalb jeder Gruppe identifiziert das Modell unregelmäßige Buchungen, indem es Muster erkennt, die erheblich vom typischen Verhalten abweichen. Die Ergebnisse dieser Analysen helfen dabei, diejenigen Teilmengen und Transaktionen zu priorisieren, die einer weiteren Prüfung durch Audit-Teams bedürfen, die Visualisierungstools zur Unterstützung ihrer Überprüfung einsetzen. Dieser Ansatz legt besonderen Wert auf Interpretierbarkeit und ermöglicht den Prüfer*innen, die Art der markierten Einträge nachzuvollziehen, anstatt sich auf komplexe oder undurchsichtige Modelle zu verlassen.

Teilbereiche der Daten werden systematisch eingestuft, um die bedeutendsten Auffälligkeiten hervorzuheben. Dies ermöglicht es dem Prüfungspersonal, Ressourcen effektiver und effizienter zuzuweisen, indem der Fokus auf Einträge gelegt wird, die als mit einem höheren Risiko behaftet eingestuft werden.

Die visuelle Darstellung spielt eine entscheidende Rolle für die Funktionalität des Tools. Streudiagramme, die nach Finanzjahren eingefärbt sind, unterstützen Prüfer*innen dabei, auffällige Trends oder Häufungen von Anomalien zu erkennen. Dies ermöglicht ihnen festzustellen, ob Unregelmäßigkeiten bestimmten Zeiträumen im Jahr zuzuordnen sind oder ob sie aus der Konsolidierung von Buchungssätzen resultieren.

Dieser datengestützte Ansatz ermöglicht eine präzisere Identifizierung von Risikobereichen und unterstützt einen gezielteren und wirksameren Prüfungsprozess.

Quelle: Auditor-General's Office (AGO) 2025 von Singapur:

KI zur Klassifizierung in Rechnungsprüfung und Aufsicht

Ähnlich wie bei der Verwendung zur Anomalieerkennung spielen Klassifikationsmodelle eine entscheidende Rolle bei der Erkennung von falschen Angaben, Betrug oder Klassifizierungsfehlern in Buchhaltungsdaten sowie bei der Kategorisierung anderer prüfungsrelevanter Dokumente und Berichte. Indem sie Datenpunkte oder Dokumente auf der Grundlage von im Datensatz erlernten Mustern vordefinierten Kategorien zuordnen, unterstützen diese Modelle Prüfer*innen dabei, Fälle oder Transaktionen zu priorisieren, die einer weiteren Überprüfung bedürfen.

Die AGO von Singapur nutzte früher einen festen regelbasierten Ansatz zur Erkennung von Anomalien in Buchungssätzen, wodurch möglicherweise Fehlklassifizierungen übersehen wurden, die den Prüfungsteams nicht bekannt waren. Um dieses Problem zu lösen, wurde ein überwachtes Machine-Learning-Modell implementiert, um potenzielle Fehlklassifizierungen zu erkennen. Dieses Modell nutzt historische Daten, einschließlich Freitextbeschreibungen, um die Kontenklassifizierung für Buchungen des aktuellen Berichtszeitraums vorherzusagen. Ein automatisierter Bewertungsschritt stuft anschließend Buchungen nach ihrer Wahrscheinlichkeit einer Fehlklassifizierung ein und unterstützt Prüfer*innen dabei, Transaktionen mit höherem Risiko für eine vertiefte Überprüfung zu priorisieren. Dieses Verfahren verbessert die statischen Regeln, indem es komplexere und bislang unbekannte Muster erkennt.

Der norwegische Riksrevisjonen hat Klassifikationsmodelle auch außerhalb von Finanzprüfungen eingesetzt. Beispielsweise wurde ein ML-Modell entwickelt, um strafrechtliche Fälle im Zusammenhang mit Informations- und Kommunikationstechnologie (IKT) aus einer großen Anzahl gemeldeter Vorfälle zu klassifizieren und zu ermitteln, welche Fälle IKT-Straftaten betreffen. Dies zeigt, wie Klassifizierung auch in den Bereichen Strafverfolgung und Aufsicht Anwendung findet. Weitere Informationen finden Sie in Kasten 2.

In einem Projekt des Controladoria-Geral da União (CGU) von Brasilien verwendet das Tool Malha Fina ein binäres Klassifikationsmodell, um die Rechenschaftsberichte der Nationalen Filmagentur (ANCINE) automatisch zu bewerten. Dabei wird festgestellt, ob die Berichte Anzeichen finanzieller Unregelmäßigkeiten aufweisen oder nicht, und es werden Empfehlungen für eine weitere persönliche Überprüfung abgegeben, wenn potenzielle Probleme erkannt werden. Im Verlauf des Projekts wurden 3.627 Rechenschaftsberichte analysiert, was einer Steigerung der Analysekapazität von ANCINE um 150% entspricht. Das Tool Malha Fina hat Einsparungen in Höhe von 803 Mio. BRL (dies entspricht 129 Mio. EUR) bei den Betriebskosten erzielt und zu einer Verkürzung der Analysezeit um 29 Jahre geführt (CGU, 2025^[31]).

Insgesamt ermöglichen Klassifizierungsmodelle Prüfungsinstitutionen, große Datenmengen systematisch und effizient über verschiedene Bereiche hinweg zu verwalten. Sie unterstützen eine gezielte Prüfungsplanung, die frühzeitige Erkennung von Unregelmäßigkeiten und fundierte Entscheidungsprozesse.

Kasten 2. Norwegischer Rechnungshof Riksrevisjonen: ML zur Identifizierung polizeilicher Maßnahmen im Zusammenhang mit IKT-Kriminalität

Ein ML-Modell wurde vom norwegischen Riksrevisjonen entwickelt, um zu ermitteln, welche gemeldeten Strafsachen mit IKT-bezogenen Straftaten in Verbindung stehen. Das Hauptziel bestand darin, das Ausmaß und die Arten von IKT-Straftaten innerhalb einer größeren Zahl gemeldeter Vorfälle zu bestimmen. Mit über 300.000 Fällen im Textformat verfolgte diese Initiative das Ziel, jeden einzelnen Fall als entweder IKT-bezogen oder nicht IKT-bezogen zu klassifizieren.

Um das Modell zu trainieren, markierten Prüfer*innen 1.072 Strafanzeigen manuell und erstellten so einen zuverlässigen Datensatz für das überwachte Lernen. Zur Analyse der Falldarstellungen wurden Techniken der Verarbeitung natürlicher Sprache (Natural Language Processing, NLP) eingesetzt, wobei der Schwerpunkt auf der Erkennung von Begriffen wie „Internet“, „Hacking“ und „Computer“ lag, um dem Modell das Erkennen relevanter Muster zu erleichtern. Es wurden mehrere Algorithmen getestet: Naive Bayes erzielte unbefriedigende Ergebnisse, während Random Forest und Gradient Boosting eine deutliche Überanpassung zeigten, wobei das Modell bei den Trainingsdaten gute Leistungen erbrachte, aber bei unbekannten Daten schlecht abschnitt. Letztlich wurde eine Support Vector Machine (SVM) ausgewählt, da sie eine überlegene Ausgewogenheit zwischen Leistungsfähigkeit und Generalisierbarkeit zeigte.

Die SVM kategorisierte erfolgreich 21.500 von insgesamt 334.544 Fällen als IKT-bezogen. Dies ergab eine deutlich höhere Anzahl von IKT-Kriminalitätsvorfällen als ursprünglich in den polizeilichen Systemen erfasst, insbesondere in den Kategorien der Finanzkriminalität. Die Diskrepanz verdeutlichte den Mehrwert des Einsatzes von ML-Tools, um verborgene Trends in umfangreichen Textdatensätzen aufzudecken.

Obwohl dies kein Prüfungsziel für den norwegische Riksrevisjonen war, könnte dieses Modell auch als praktisches Hilfsmittel für die Polizei dienen, um eine fundiertere Ressourcenzuweisung und strategische Planung zu ermöglichen. Die Prüfungsergebnisse wurden von den Strafverfolgungsbehörden positiv aufgenommen und verdeutlichen den Mehrwert datenbasierter Entscheidungsunterstützung. Wichtig ist, dass das Projekt vollständig intern durchgeführt wurde, ohne auf externe Berater*innen zurückzugreifen, was Kosteneffizienz und vollständige Kontrolle über den Prozess gewährleistete.

Quelle: Riksrevisjonen 2025.

Prädiktive und präventive Rechnungsprüfungsmodelle

Eine zunehmende Anzahl von Aufsichtsbehörden prüft den Einsatz von KI- und ML-Modellen, um von reaktiven Prüfungen und Aufsicht – die darauf ausgerichtet sind, Probleme erst im Nachhinein zu identifizieren – zu eher präventiven Ansätzen überzugehen. Prädiktive Modelle sind darauf ausgelegt, Risikoevents vorherzusehen oder Frühwarnmuster zu erkennen, sodass Prüfer*innen ihre Aufmerksamkeit auf Bereiche mit einem höheren Potenzial für Regelverstöße oder Systemausfälle richten können. Auf diese Weise ermöglichen diese Tools ein früheres Eingreifen und eine strategischere Prüfungsplanung, sodass Aufsichtsbehörden ihre Ressourcen auf Bereiche konzentrieren können, in denen Probleme mit hoher Wahrscheinlichkeit auftreten werden.

Diese Tools basieren in der Regel auf Klassifikationsmodellen oder prädiktiven Algorithmen, die große Mengen strukturierter und/oder unstrukturierter Daten analysieren[†]. Datenquellen können beispielsweise

Beschaffungsunterlagen, Planungsdokumente, Finanzdaten, Informationen zu wirtschaftlichen Eigentümern, Vertragsdatenbanken sowie Daten zu Infrastrukturprojekten umfassen. Die Modelle liefern priorisierte Risikobewertungen, die Prüfer*innen sowie Aufsichtsbehörden dabei helfen, ihre Analysen und Kontrollen gezielter und wirkungsvoller auszurichten. Ein detailliertes Beispiel für ein solches System, das vom italienischen Rechnungshof (*Corte dei conti*) verwendet wird, ist in Kasten 3 vorgestellt. Während KI die Geschwindigkeit und den Umfang der Risikoanalyse erhöht, bleiben die Prüfer*innen bislang für die endgültigen Entscheidungen und Validierungen verantwortlich.

Ein weiteres Anwendungsbeispiel ist der Einsatz von KI beim CGU in Brasilien, wo Prüfungsstellen Tools entwickelt haben, um die präventive Überprüfung von Beschaffungsprozessen zu unterstützen. Ein solches Tool, das im Rahmen eines Pilotprojekts entwickelt wurde, nutzt LLMs, um die Durchführbarkeit von IT-Verträgen sowie deren Konformität mit gesetzlichen Rahmenbedingungen zu bewerten. Frühe Ergebnisse zeigen, dass KI-gestützte Bewertungen von Dutzenden von Beschaffungsprozessen deutlich schneller und kostengünstiger durchgeführt werden können.

In einem weiteren Beispiel in Italien wurden ebenfalls prädiktive Modelle entwickelt, um die frühzeitige Erkennung von Risiken bei öffentlichen Vergabe- und Infrastrukturprojekten zu unterstützen. Mithilfe von Daten der National Anti-Corruption Authority (ANAC) und anderer öffentlicher Vergabedatenbanken erstellt das System Risikoindikatoren, um die Überwachung von Instandhaltungs- und Infrastrukturprojekten zu priorisieren. Dieser Ansatz unterstützt die Echtzeit- bzw. kontinuierliche Prüfung, indem er Teilprojekte mit erhöhtem Risikoniveau hervorhebt und es den Prüfungsinstitutionen ermöglicht, Ressourcen dort zu bündeln, wo sie am dringendsten benötigt werden.

Obwohl sie sich noch in der Entwicklung befinden, verdeutlichen prädiktive und präventive Prüfungsmodelle das Potenzial von KI, das proaktive Risikomanagement zu stärken, die Wahrscheinlichkeit zukünftiger Fehler zu verringern und die Effizienz der Prüfungsplanung zu verbessern. Es ist jedoch wichtig, sich bewusst zu sein, dass eine übermäßige Abhängigkeit von prädiktiven Modellen dazu führen kann, neu auftretende Risikofaktoren zu übersehen, die in den Trainingsdaten noch nicht abgebildet sind, was die Genauigkeit eines Modells im Laufe der Zeit potenziell verringern kann.

Kasten 3. Italiens Corte dei conti: Vorhersage von Gemeinden in finanziellen Schwierigkeiten

Ein Machine-Learning (ML)-Ansatz, erweitert durch Fachwissen

Die finanzielle Stabilität der Kommunen ist ein zentrales Anliegen der öffentlichen Verwaltung, da finanzielle Notlagen zur Unfähigkeit führen können, essenzielle Dienstleistungen bereitzustellen, und dadurch das Wohlergehen der Gemeinschaften beeinträchtigen. In Italien sind lokale Regierungen gesetzlich verpflichtet, ihren Betrieb auch in finanziellen Krisenzeiten aufrechtzuerhalten, sodass eine frühzeitige Erkennung und Intervention unerlässlich ist. Um dieser Herausforderung zu begegnen, hat der italienische Corte dei conti RoMoLo entwickelt, ein experimentelles ML-Modell, das darauf ausgelegt ist, finanzielle Notlagen in Gemeinden vorherzusagen.

Die Entwicklung von RoMoLo umfasste die Erstellung eines umfassenden Datensatzes, der 7.094 Gemeinden über einen Zeitraum von sieben Jahren (2016–2022) abdeckt. Der Datensatz umfasst 37 Attribute, die aus heterogenen Datenquellen abgeleitet wurden, darunter Jahresabschlüsse, demografische Informationen und institutionelle Indikatoren. Die Daten wurden vor dem Training des Modells neu gewichtet: Da nur ein geringer Prozentsatz der Gemeinden von finanziellen Schwierigkeiten betroffen ist, wurde deren Gewichtung in den Daten angepasst, um Verzerrungen zu vermeiden.

Die in den Datensatz aufgenommenen Attribute wurden aufgrund ihrer Relevanz für die finanzielle Solidität ausgewählt und basieren auf der Expertise von Fachkräften aus der Rechnungsprüfung. Sie

umfassen verschiedene Aspekte des kommunalen Handelns, wie demografische Merkmale, geografische Lage, Investitionsniveau, finanzielle Autonomie, Verschuldung, Ausgabenrigidität und Praktiken des Schuldenmanagements.

Eine „Insolvenzrisiko“-Variable wird vom prädiktiven Modell erzeugt; diese Funktion ermöglicht die Einstufung jeder Gemeinde in eine Risikokategorie auf Grundlage des Auftretens schwerwiegender finanzieller Ungleichgewichte. Die fünf Risikostufen sind:

- Geringes Risiko: hat weder Vorzeichen einer finanziellen Notlage noch eine finanzielle Notlage erlebt.
- Mittleres-niedriges Risiko: hat mindestens einen Fall von Vorbelastung erlebt.
- Mittleres Risiko: hat mindestens einen Fall finanzieller Notlage erlebt.
- Mittleres bis hohes Risiko: hat sowohl eine Vorbelastung als auch anschließend eine finanzielle Notlage erfahren.
- Hohes Risiko: hat mehrere Fälle finanzieller Notlage erlebt.

Zur Unterstützung dieser Klassifizierung wurden mehrere ML-Modelle getestet, darunter Gaussian Naive Bayes, Logistic Regression, Random Forest, Support Vector Machine, Gradient Boosting sowie ein benutzerdefinierter Ensemble-Ansatz. Logistic Regression wurde aufgrund ihrer hohen Leistungsfähigkeit und Robustheit bei der Erfassung des Zusammenhangs zwischen Prädiktoren und den definierten Risikokategorien ausgewählt. RoMoLo erreichte eine Sensitivität („Recall“) von 1, was bedeutet, dass alle Gemeinden in finanziellen Schwierigkeiten korrekt identifiziert wurden, ohne dass es zu Fehlklassifikationen negativer Fälle kam. Das Modell erzeugte jedoch auch eine beträchtliche Anzahl von falsch-positiven Ergebnissen, was darauf hindeutet, dass das Modell zwar sehr effektiv bei der Erkennung von Gemeinden in finanzieller Not ist, aber auch einige Gemeinden markieren kann, die tatsächlich nicht betroffen sind, und daher eine weitere Überprüfung und Validierung erforderlich machen.

Das Hauptziel von RoMoLo besteht darin, dem Rechnungshof als Frühwarnsystem für Gemeinden mit Risiko einer finanziellen Notlage zu dienen. Durch die Bereitstellung zeitnaher Prognosen ermöglicht das Modell Entscheidungsträger*innen und Prüfer*innen proaktiv einzugreifen und Korrekturmaßnahmen zu ergreifen, bevor eine Krise eskaliert.

RoMoLo stellt einen bedeutenden Fortschritt bei der Anwendung von ML im Bereich der öffentlichen Finanzprüfung dar. Die Ergebnisse dieser Studie zeigen, dass der Einsatz von ML-Modellen in Kombination mit dem Fachwissen von Expert*innen des Rechnungs-Rechtswesens die frühzeitige Erkennung finanzieller Schwierigkeiten unterstützen kann, was zu besseren Ergebnissen für die Gemeinschaften führt.

Quelle: Corte dei conti, 2025.

Suche, Abruf und Wissensmanagement

Prüfungsinstitutionen setzen zunehmend KI ein, um die Dokumentensuche, das semantische Verständnis sowie das Auffinden prüfungsrelevanten Wissens oder von Rechtsvorschriften zu verbessern. Wissensmanagement bleibt eine zeitaufwändige Tätigkeit, die häufig von Prüfer*innen verlangt, große Mengen an textbasierten Materialien, einschließlich geprüfter Daten, Standards, Richtlinien und Vorschriften, zu durchsuchen. KI-Tools in diesem Bereich fungieren als „Prüfungshelfer“, indem sie den Zugang zu relevanten Informationen durch automatisierte Dokumentenerkennung, semantische Suchtechniken[†] und Frage-Antwort-Funktionen über umfangreiche Inhaltsbestände hinweg erleichtern.

Selbst wenn die Ergebnisse nicht vollkommen präzise sind, übertreffen diese Systeme in der Regel manuelle Methoden, die auf exakter Wort- oder Phrasensuche basieren, indem sie relevante Inhalte schneller und konsistenter auffinden. Durch das Erfassen der semantischen Bedeutung von Nutzeranfragen können sie mit Synonymen oder unterschiedlichen Formulierungen sowie komplexen Suchausdrücken umgehen und Aufgaben wie Zusammenfassungen und Themenextraktion automatisieren. Dadurch unterstützen sie Prüfer*innen bei einer schnelleren Analyse und Entscheidungsfindung. Dies stellt eine Anwendung von KI mit geringerem Risiko dar, die die Effizienz steigert und höhere Akzeptanzraten innerhalb von Prüfungsinstitutionen unterstützt.

Über die Verbesserung der Informationsbeschaffung hinaus kann KI dazu beitragen, die Art und Weise zu transformieren, wie Prüfungsinstitutionen ihr gesammeltes Wissen organisieren, interpretieren und anwenden. Fortgeschrittene Sprach- und semantische Modelle können Muster und Zusammenhänge in früheren Prüfungen, regulatorischen Aktualisierungen und aufkommenden Risiken erkennen und so Prüfer*innen dabei unterstützen, Erkenntnisse zu gewinnen, die andernfalls in verstreuter Dokumentation verborgen bleiben könnten. In Kombination mit KI-gestützten Zusammenfassungs- und Klassifizierungswerkzeugen können sie dazu beitragen, aktuelle Prüfmethodologien, Leitfäden und Schulungsressourcen auf dem neuesten Stand zu halten. Diese Fähigkeit stärkt das institutionelle Lernen, fördert eine größere Einheitlichkeit zwischen den Prüfungsteams und verbessert letztlich die Qualität, Agilität und Reaktionsfähigkeit der öffentlichen Rechnungsprüfung.

Prüfungsinstitutionen verfolgen bei der Implementierung dieser Systeme einen zweigleisigen Ansatz: Einige nutzen handelsübliche Standardlösungen, während andere eigene Modelle entwickeln, die auf auditspezifische Terminologie und Inhalte trainiert sind. Zu den Schlüsseltechnologien zählen semantische Suchmaschinen, Retrieval-Augmented Generation (RAG)-Pipelines, Vektordatenbanken und domainspezifisch angepasste LLMs.[†]

In Deutschland hat der Bundesrechnungshof (BRH) ein kommerzielles semantisches Suchsystem implementiert, das an seine Anforderungen angepasst wurde. Dieses System soll Prüfer*innen dabei unterstützen, große Mengen von Dokumenten in unterschiedlichen Formaten zu durchsuchen. Es vereint Volltextsuche, Metadatenanreicherung und Filterfunktionen. Das System ermöglicht außerdem die Nutzung von Python-basierten Skripten zur Unterstützung fortgeschrittener Suchszenarien. Es ist seit über einem Jahr aktiv im Einsatz und trägt Berichten zufolge erheblich dazu bei, den Zeitaufwand für die Informationsbeschaffung zu reduzieren.

Über die reine Dokumentenrecherche hinaus werden KI-gestützte Chatbots und virtuelle Assistenten zunehmend eingesetzt, um Rechnungsprüfer*innen intern bei der Bereitstellung von Anleitungen, der Beantwortung von Anfragen, zur Methodik, zu Prüfungsstandards sowie bei technischen Fragestellungen zu unterstützen. Diese interaktiven Tools erleichtern die Kommunikation, indem sie es Prüfer*innen sowie Mitarbeiter*innen ermöglichen, Fragen in natürlicher Sprache zu stellen und kontextbezogene Antworten zu erhalten. Dadurch werden Arbeitsabläufe effizienter gestaltet und die Einarbeitung verbessert. Beispiele für solche Anwendungen finden sich in verschiedenen Institutionen:

- Der Europäische Rechnungshof (EuRH) erwägt den Einsatz eines internen Chatbots zur Unterstützung der Prüfmethodik, der den Prüfer*innen einen leichteren Zugang zu methodischen Informationen, Checklisten, Qualitätsstandards und Leitlinien bietet. Sie testen außerdem die Integration von Microsoft 365 Copilot, um die Produktivität zu steigern, indem Informationen aus Tools wie Outlook, Teams oder SharePoint extrahiert werden, wobei stets die Anforderungen an den Datenschutz eingehalten werden.
- Im Contraloría General de la República (CGR) von Costa Rica nutzt das Prüfungsamt LLM-gestützte Notebooks als interne Assistenten, die es den Mitarbeiter*innen ermöglichen, Datenbanken und institutionelle Verfahren im Zusammenhang mit Personalwesen, IT-Management und Beschaffung abzufragen. Diese Anwendung hat das interne Wissensmanagement und die interne Beratung verbessert und zugleich die

Einarbeitungsprozesse neuer Mitarbeiter*innen beschleunigt, indem sie einen effizienten Zugang zu präzisen und aktuellen institutionellen Informationen ermöglicht. Obwohl diese Anwendungen nicht unmittelbar Teil der Prüfungsarbeit sind, tragen sie zu einer effizienteren Ressourcenverwaltung bei, sparen den Prüfer*innen Zeit und fördern die Standardisierung und Verbreitung von institutionellem Wissen, was letztlich qualitativ hochwertigere und einheitlichere Prüfungsprozesse unterstützt.

- Weitere Beispiele werden für Brasilien in Kasten 4 und Frankreich in Kasten 6 vorgestellt.

Da Prüfungsinstitutionen mit umfangreichen und häufig fragmentierten Wissensbeständen konfrontiert sind, bieten KI-gestützte Tools für Suche, Informationsabruf, Wissensmanagement und interaktive Prüfungsunterstützung einen pragmatischen Einstieg in die Einführung von KI. Auch wenn weiterhin Herausforderungen bestehen, darunter die Notwendigkeit qualitätsgesicherter Dokumentenkorpora, sorgfältiges Prompt-Design, Datensicherheit und Nutzervertrauen, erweisen sich solche Tools als wirkungsvoll bei der Steigerung der internen Effizienz und der Unterstützung fundierterer Prüfungsarbeiten.

Kasten 4. Controladoria-Geral da União (CGU) von Brasilien: AuditPesquisa

Beim CGU in Brasilien wurde ein hochentwickeltes, chatbasiertes Tool eingeführt, um die Prüfung und Auswertung von Prüfungsberichten zu verbessern. Dieses Tool wurde speziell entwickelt, um den Zugang zu einem umfangreichen Bestand an Prüfungsinformationen zu erleichtern, indem es den Nutzer*innen (z. B. Prüfer*innen) ermöglicht, mithilfe natürlicher Sprache mit dem System zu interagieren und so eine nahtlose und intuitive Nutzung zu gewährleisten. Anstatt sich manuell durch umfangreiche, komplexe Dokumente zu arbeiten, können die Nutzer*innen Fragen stellen, wie sie es in einem Gespräch tun würden, und das System liefert klare, relevante Antworten, die direkt aus den Berichten stammen. Dadurch wird es einfacher, zentrale Erkenntnisse zusammenzufassen, Risiken zu identifizieren, Muster über mehrere Prüfungen hinweg zu erkennen und wichtige Themen, die in früheren Bewertungen hervorgehoben wurden, gezielt zu bestimmen.

Die Funktionsweise des Tools basiert auf dem Retrieval-Augmented Generation (RAG)[†]-Framework, das in einem vierstufigen Prozess arbeitet:

1. Nutzer*in stellt eine Anfrage.
2. Das System tritt in eine Abrufphase ein, in der es seine Datenbank mit Prüfungsberichten (im Folgenden erläutert) durchsucht, um die relevantesten Abschnitte zu identifizieren.
3. Es folgt eine Analysephase, in der die OpenAI API LLM-Technologie das abgerufene Material interpretiert.
4. Eine gut kontextualisierte Antwort, die auf die ursprüngliche Anfrage zugeschnitten ist, wird erstellt.

Der Einsatz fortschrittlicher KI-Technologien wie ChatGPT bildet die Grundlage für die Fähigkeit des Systems, kohärente und präzise Ergebnisse zu erzeugen.

Dieses Tool ist nun im Einsatz; seine Entwicklung verlief jedoch nicht ohne Herausforderungen. Die größte Herausforderung bestand in der Erstellung und Standardisierung der Datenbank für Prüfungsberichte. Insgesamt konnten 3.189 öffentlich verfügbare Prüfungsberichte in den Anfangsdatensatz aufgenommen werden. Es war erforderlich, sie in ihre Kernelemente zu unterteilen – typischerweise Einleitung, Prüfungsergebnisse, Empfehlungen und Fazit –, da diese häufig im Mittelpunkt von Nutzeranfragen stehen. Im Rahmen der Datenaufbereitung wurden zudem nicht wesentliche Elemente wie Tabellen, Grafiken, Abbildungen und zusammenhangslose Aussagen

entfernt, da diese das System während der Analyse- und Generierungsphasen andernfalls verwirren oder in die Irre führen könnten.

Neben der Inhaltsextraktion erfasste das System auch eine breite Palette von Metadaten zu jedem Bericht, einschließlich Informationen über die geprüfte Stelle, die leitende oder überwachende Einheit, den geografischen Standort, die Art der Dienstleistung sowie die Kategorie der geprüften Tätigkeit. Diese Metadaten spielen eine entscheidende Rolle bei der Erhöhung der Präzision des Tools, indem sie den Einsatz kontextbezogener Filter ermöglichen und so die Relevanz und Genauigkeit der bereitgestellten Antworten steigern. Nutzer*innen können ihre Suche nach Abteilung, Regierungsstelle, Region, Auditkategorie oder Dienstleistungsart eingrenzen, was eine gezielte und effiziente Informationsbeschaffung ermöglicht. Diese Funktion verwandelt einen früher arbeitsintensiven Prozess in eine effiziente Aufgabe, verkürzt die Zeit für die Auswertung zentraler Ergebnisse und verbessert den Entscheidungsprozess. Letztlich stellt diese Innovation einen bedeutenden Fortschritt bei der Anwendung von KI im Prüfungswesen dar und fördert Transparenz, operative Effizienz sowie strategische Bewertung im gesamten Sektor.

Quelle: Controladoria-Geral da União (CGU), Brasilien, 2025.

Intelligente Dokumentenverarbeitung (IDP) und Datenextraktion

Intelligente Dokumentenverarbeitung (IDP) bezeichnet den Einsatz von KI, um automatisch strukturierte Daten aus unstrukturierten oder semi-strukturierten Quellen wie Portable Document Formats (PDFs), gescannten Dokumenten, Bildern und handschriftlichen Formularen zu extrahieren. Diese Systeme kombinieren Technologien wie optische Zeichenerkennung (Optical character recognition, OCR), Natural Language Processing (NLP) und LLMs, um Dokumente zu klassifizieren, zentrale Entitäten zu identifizieren und Freitext in nutzbare, strukturierte Formate umzuwandeln.

Im Bereich der öffentlichen Rechnungsprüfung gewinnen IDP-Tools zunehmend an Bedeutung, da sie dazu beitragen, umfangreiche und wiederkehrende Aufgaben wie die Überprüfung von Belegen, Rechnungen, Verträgen und Spesenabrechnungen effizienter zu gestalten. Diese Dokumente werden derzeit überwiegend manuell analysiert, häufig mithilfe von Stichprobenverfahren. IDP bietet das Potenzial, diese Prozesse in großem Umfang zu automatisieren, die Genauigkeit zu erhöhen, den Zeitaufwand für die Dokumentenbearbeitung zu verringern und Voll-Prüfungen der Daten zu ermöglichen.

Im Gegensatz zu umfassenderen Wissensmanagementsystemen sind IDP-Tools für die gezielte Extraktion bestimmter Felder oder Werte konzipiert, beispielsweise Namen, Daten und Beträge aus unterschiedlichen Dokumentformaten. Wenn sie in umfassendere Prüfungsabläufe integriert werden, können IDP-Systeme Verifizierungsprozesse verbessern, die frühzeitige Erkennung von Anomalien unterstützen und die administrative Belastung der Prüfungsteams verringern.

So experimentieren beispielsweise Prüfungsteams bei dem italienischen Corte dei Conti mit der automatisierten Extraktion von Daten aus PDF- und Bilddateien, um stichprobenbasierte Prüfungen öffentlicher Ausgaben auf Bevölkerungsebene zu unterstützen. Diese Maßnahmen konzentrieren sich darauf, Text- und Zahlenwerte automatisch aus Belegen zur Ausgabenrechtfertigung zu identifizieren und zu extrahieren. Während dieser Prozess derzeit manuell durchgeführt wird, besteht das Ziel darin, ihn vollständig zu automatisieren.

Kasten 5. Italienischer Corte dei conti – GiusBERTo

Eines der frühesten KI-Experimente am italienischen Corte dei conti war GiusBERTo, ein Proof of Concept, der entwickelt wurde, um die Anonymisierung sensibler personenbezogener Daten in Gerichtsunterlagen zu automatisieren. GiusBERTo wurde auf der Grundlage Tausender früherer Gerichtsentscheidungen trainiert und erkennt und schwärzt personenbezogene Daten wie Orte, Namen und Daten.

GiusBERTo wurde einer umfassenden Prüfung anhand standardisierter Anonymisierungsmetriken unterzogen, und die Ergebnisse zeigten die praktische Wirksamkeit des Modells bei der Anwendung auf bereits anonymisierte tatsächliche Urteile. Dies belegt die Fähigkeit des Tools, die Einhaltung der Datenschutzgesetze zu gewährleisten und gleichzeitig die notwendige Transparenz in öffentlichen Aufzeichnungen zu wahren.

Aufbauend auf dieser Grundlage werden weitere Anstrengungen unternommen, um die Anonymisierung zu verbessern. Mit den technologischen Fortschritten und dem zunehmenden Einfluss generativer KI wurde das Modell neu gestaltet, um modernste Funktionen zu nutzen. Dabei wurden Human-in-the-Loop-Methoden integriert, die es den Nutzer*innen ermöglichen, Ergebnisse bei Bedarf zu validieren und zu korrigieren. Darüber hinaus ist das Tool mit cloudbasierten Plattformen wie Amazon Web Services (AWS) integriert, um Skalierbarkeit und Leistung zu verbessern, sodass es nun für Tests durch Endnutzer*innen bereitsteht.

Mit der Weiterentwicklung der KI-Fähigkeiten des Rechnungshofs erweitert sich die Initiative über die grundlegende Anonymisierung hinaus hin zu fortschrittlichen Text-Mining-Techniken, die eine umfassendere Überprüfung öffentlicher Ausgaben ermöglichen. Diese Entwicklung zeigt das Engagement des Rechnungshofs, KI zur Verbesserung der Governance und Rechenschaftspflicht bei öffentlichen Ausgaben einzusetzen.

Quelle: Corte dei conti, 2025.

Andernorts wird IDP derzeit geprüft, um die Analyse von Personalakten im öffentlichen Dienst zu unterstützen. Dabei werden KI-Modelle eingesetzt, um strukturierte Daten wie Namen oder akademische Abschlüsse aus archivierten Dokumenten zu extrahieren und in maschinenlesbare Formate umzuwandeln. Dieser Prozess ermöglicht eine schnellere Verarbeitung und Integration von aus Dokumenten gewonnenen Informationen in andere analytische Systeme.

IDP stellt eine vielversprechende Perspektive für die Automatisierung von Prüfungen dar. Zu den wichtigsten Vorteilen zählen eine verbesserte Dokumentenabdeckung, ein geringerer manueller Aufwand und die Möglichkeit, Analyse-Tools konsistent auf vollständige Datensätze anzuwenden. Dennoch müssen die Institutionen weiterhin Herausforderungen wie die OCR-Qualität – die sich zwar schnell verbessert, aber noch nicht vollständig zuverlässig ist –, den Datenschutz sowie die Notwendigkeit einer domainspezifischen Feinabstimmung der Modelle bewältigen.

Generative KI für das Verfassen, Zusammenfassen und Übersetzen

Aufsichtsbehörden und öffentliche Organisationen insgesamt nutzen zunehmend LLMs, um die Texterstellung zu unterstützen oder zu automatisieren, einschließlich der Erstellung von Prüfungsdokumenten, der Zusammenfassung technischer Unterlagen, der Vorbereitung von Skripten für

die öffentliche Kommunikation und der Übersetzung von Inhalten in verschiedene Sprachen. In allen Fällen fungieren KI-Tools auch hier als Assistenzsysteme und nicht als endgültige Entscheidungsinstanzen, indem sie erste Entwürfe oder Vorschläge erstellen, die anschließend von Prüfer*innen oder Kommunikationsmitarbeiter*innen geprüft und validiert werden.

Der Reiz generativer KI liegt über die Steigerung der Produktivität hinaus in ihrer Fähigkeit, Texte zu erstellen, die im Tonfall und in der Terminologie einheitlich sind. Gleichzeitig bestehen weiterhin Bedenken hinsichtlich der Genauigkeit der generierten Ergebnisse, insbesondere in risikobehafteten Kontexten. Dies ist insbesondere bei der Übersetzung technischer Prüfungsthemen relevant, da hierbei präzise Terminologie oder sprachliche Nuancen, die für bestimmte Rechtsordnungen spezifisch sind, in unterschiedlichen Kontexten an Genauigkeit oder Bedeutung verlieren können. Die meisten Institutionen mindern diese Risiken, indem sie mehrstufige Prüfprozesse aufrechterhalten und eine eindeutige menschliche Aufsicht gewährleisten.

Beim EuRH wird Microsoft 365 Copilot von einer Nutzergruppe für verschiedene Anwendungsfälle getestet, darunter die Erstellung von Textentwürfen. Die Einführung dieses Tools wurde von einem internen Beratungsausschuss für KI begleitet, der dazu beigetragen hat, angemessene Nutzungsbedingungen festzulegen. Das Tool kann ein wertvolles Instrument sein, um die Erstellung von Prüfungsdokumentationen in der Anfangsphase zu beschleunigen. Um jedoch das Risiko von Fehlern in KI-generierten Texten zu behandeln, sind menschliche Aufsicht und Überprüfung unerlässlich, und die Nutzer*innen bleiben letztlich für die erstellten Texte verantwortlich. Die Risiken werden zusätzlich durch die strikten Qualitätskontrollen und mehrstufigen Prüfverfahren der Institution gemindert.

In Costa Rica wird generative KI eingesetzt, um die Erstellung von Drehbüchern für öffentlichkeitswirksame Kommunikationsmaterialien zu unterstützen. Mit Unterstützung von LLMs hat das Prüfungsamt vorläufige Skripte für Video-Briefings und bürgerorientierte Publikationen erstellt, die anschließend von Kommunikationsbeauftragten geprüft und validiert werden. Durch die Automatisierung eines Teils des Prozesses der Drehbucheerstellung hat die Institution die Klarheit, Konsistenz und Agilität ihrer öffentlichen Kommunikation gestärkt.

Sowohl der CGR Costa Ricas als auch der EuRH prüfen zudem den Einsatz generativer KI zur Zusammenfassung umfangreicher Dokumentensammlungen. In Costa Rica unterstützen LLMs dabei, umfangreiche technische Dokumente zusammenzufassen und erste Entwürfe von Arbeitspapieren sowie Prüfungsergebnissen zu erstellen, die anschließend von Prüfer*innen überprüft und validiert werden.

Ein weiteres Beispiel für den Einsatz von LLMs durch den Rechnungshof Italiens zur Anonymisierung amtlicher Dokumente wird in Kasten 5 vorgestellt.

Generative KI spielt auch eine zunehmend wichtige Rolle bei der Übersetzung und Überprüfung von Dokumenten. Bei der multinationalen Rechnungsprüfungsgesellschaft werden KI-Modelle eingesetzt, um Prüfungsdokumentationen in mehrere Sprachen zu übersetzen – eine bedeutende operative Herausforderung im internationalen Prüfungswesen. Alle übersetzten Ergebnisse werden anschließend von internen Fachexpert*innen geprüft und qualitätsgesichert, um Genauigkeit und Angemessenheit sicherzustellen. Das Unternehmen hat zudem ein internes generatives KI-basiertes Tool entwickelt, das informell als „Review Agent“ bezeichnet wird. Ursprünglich als eine Persona konzipiert, die Prüfer*innen simuliert, unterstützt dieses Tool die Mitarbeiter*innen bei der Überprüfung von Prüfungsberichten, indem es unklare Abschnitte kennzeichnet, Widersprüche identifiziert oder Verbesserungsvorschläge macht. Die Nutzung des Tools ist optional, und obwohl es keine komplexen Auslegungsfragen wie die Anwendung spezifischer gesetzlicher oder regulatorischer Rahmenbedingungen lösen kann, wird es als hilfreich angesehen, um die Konsistenz und Lesbarkeit von Prüfungsunterlagen zu verbessern.

Eine weitere Anwendung von generativer KI zur Unterstützung beim Verfassen und Zusammenfassen wird Kasten 6 vorgestellt.

Kasten 6. Französischer Cour des comptes – ChatJF: Entwicklung eines souveränen generativen KI-Assistenten

Als Reaktion auf das zunehmende institutionelle Interesse an KI und den Bedarf an sicheren, souveränen digitalen Tools wurde vom französischen Cour des comptes ein hausinterner Prototyp namens ChatJF entwickelt. Ziel ist es, die praktischen Anwendungen von LLMs in Finanzgerichtsbarkeiten zu untersuchen, mit besonderem Fokus auf Dokumentenzusammenfassungen, Übersetzungen, Paraphrasierungen, Datenanalysen und Recherchen. Die Initiative verfolgte zudem das Ziel, die vollständige Einhaltung gesetzlicher und regulatorischer Vorgaben sicherzustellen, insbesondere im Hinblick auf die Datenschutz-Grundverordnung (DSGVO) und die berufliche Schweigepflicht.

Erste Anwendungsfälle wurden in gemeinsamen Workshops mit Mitarbeiter*innen ermittelt und umfassten die Zusammenfassung von Dokumenten, Unterstützung beim Verfassen und Übersetzen, Dokumentenrecherche, Analyse von Datentabellen sowie Transkription. Obwohl diese Anwendungsfälle identifiziert wurden, sind nicht alle vollständig umgesetzt; ihre Weiterentwicklung erfolgt entsprechend den festgelegten Prioritäten. Anschließend wurden fortschrittlichere Anwendungen in die Liste aufgenommen, darunter Generierung von Inhalten, Erstellung von Infografiken und Unterstützung bei der Auditvorbereitung. Mit zunehmender Nutzung begann die Plattform, Konversationsagenten und RAG zu unterstützen, sodass Nutzer*innen Dokumente hochladen und direkt abfragen können.

ChatJF wurde mithilfe von Open-Source-Technologien entwickelt, um vollständige Kontrolle über Daten und Systemarchitektur zu gewährleisten. Seine drei zentralen Komponenten sind ein Open-Source-LLM-Manager, OpenWebUI – eine lokal gehostete Weboberfläche nach dem Vorbild von ChatGPT – sowie ein Add-in für Microsoft Word, das eine nahtlose Integration ermöglicht.

Das System integriert derzeit mehrere LLMs, darunter beispielsweise LLaMA 3 von Meta und ein Modell von Mistral, wobei je nach Aufgabe ein Wechsel zwischen den Modellen möglich ist. Um Leistung und Anpassungsfähigkeit zu gewährleisten, erfolgt die Modellausführung auf Grafikprozessoren (GPUs), die von einem französischen Cloud-Anbieter gehostet werden, während die Schnittstelle und die Berechnung von Embeddings intern auf virtuellen Maschinen abgewickelt werden.

Eine bemerkenswerte Neuerung ist die Integration von ChatJF in Microsoft Word, wodurch Funktionen wie Zusammenfassung, Umformulierung und Übersetzung ermöglicht werden. Dies wurde intern mithilfe von TypeScript, HTML und CSS entwickelt und verbindet sich über individuelle API-Schlüssel mit der Hauptplattform. Das Plugin unterstützt Office 2019 und ist auf die übergeordneten Produktivitäts-Tools der Organisation abgestimmt, wodurch die Einführung von KI unter den Nutzer*innen erleichtert wird.

Das System befindet sich derzeit im Prototypenstadium, wird jedoch aktiv genutzt: Innerhalb von 30 Tagen wurden etwa 300 eindeutige Nutzer*innen verzeichnet, davon rund 100 regelmäßige Nutzer*innen. Etwa 20% der Mitarbeiter*innen der Organisation haben das Tool genutzt. Während die Plattform noch getestet wird, automatisieren die Nutzer*innen bereits die Transkription, die Erstellung von Inhalten und die strukturierte Dokumentenanalyse. Die Einführung erforderte umfangreiche Schulungs- und Kommunikationsmaßnahmen, um Erwartungen zu steuern und eine sachgemäße Nutzung zu fördern.

Sicherheit und Datensouveränität standen im Mittelpunkt der Projektgestaltung. Alle Komponenten werden auf internen Servern oder bei zertifizierten externen Anbietern gehostet. Keine Daten verlassen den institutionellen Bereich, sodass gewährleistet ist, dass sensible Informationen, die der gesetzlichen

Geheimhaltung unterliegen, unter vollständiger Einhaltung der rechtlichen Rahmenbedingungen verarbeitet werden.

Der Plan des französischen Cour des comptes umfasst den Übergang vom Prototyp zur Produktion, die Vervielfachung der Infrastruktur, die Integration von Benutzerunterstützungssystemen sowie die Erweiterung der Microsoft Office-Integration. Die explorative Arbeit wird mit Spracherkennung und automatisierten Visualisierungstools fortgesetzt. Mit ChatJF zeigt der französische Cour des comptes einen skalierbaren Ansatz für generative KI, indem diese in zentrale Arbeitsabläufe integriert wird und gleichzeitig die Kontrolle über Daten und Infrastruktur gewahrt bleibt.

Quelle: Cour des comptes, 2025.

Mit der fortschreitenden Entwicklung generativer KI erhalten Prüfungs- und Aufsichtsbehörden zunehmend wirkungsvolle Unterstützung bei sprachbasierten Aufgaben mit hohem Volumen. Diese Tools beschleunigen interne Arbeitsabläufe, verkürzen den Zeitaufwand für das Verfassen und Zusammenfassen von Texten und verbessern die Verständlichkeit in der externen Kommunikation. Dennoch verfolgen die Institutionen weiterhin einen abgewogenen Ansatz, wobei der Schwerpunkt auf Validierung, Transparenz und der Gewährleistung liegt, dass KI eindeutig eine unterstützende Funktion behält.

Visuelle und räumliche Datenanalyse

Weniger verbreitet, aber vielversprechend sind KI-Anwendungen zur Analyse von visuellen und räumlichen Daten, einschließlich Bildern, Satellitenaufnahmen und von Drohnen erfassten Inhalten, die Leistungs- und Compliance-Prüfungen unterstützen können. Diese Tools können neue Formen von Nachweisen für die Bewertung physischer Infrastrukturen, die Überwachung der Flächennutzung oder die Überprüfung der Umsetzung von politischen Maßnahmen vor Ort liefern. Ihre Einführung wird jedoch durch mehrere Faktoren eingeschränkt, darunter rechtliche Unsicherheiten hinsichtlich der Nutzung von Internet-of-Things-(IoT)-Geräten wie Drohnen, Einschränkungen bei der Datenqualität sowie die hohen Kosten, die mit dem Aufbau von Kompetenzen verbunden sind.

Diese Tools sind insbesondere dort von Bedeutung, wo herkömmliche Prüfungsverfahren auf physische oder logistische Einschränkungen stoßen. KI kann die automatisierte Auswertung visueller Daten unterstützen, indem Methoden wie Computer Vision und Deep-Learning-Modelle, etwa Convolutional Neural Networks (CNNs)[†], eingesetzt werden, um Muster zu erkennen, Objekte zu klassifizieren oder die Einhaltung von Vorschriften aus der Ferne zu bewerten. Obwohl das Potenzial erheblich ist, befinden sich diese Ansätze häufig noch in der experimentellen Phase.

In Costa Rica verfolgt eine Forschungsinitiative das Ziel, ein lokal produziertes, kostengünstiges IoT-Kamerasystem zur Überwachung öffentlicher Bauprojekte zu entwickeln. Das System integriert künstliche Intelligenz, um Fortschrittmuster im Bauwesen zu erkennen und so eine frühzeitige Erkennung von Verzögerungen oder Unregelmäßigkeiten durch öffentliche Prüfer*innen zu unterstützen. Diese Initiative stellt ein vielversprechendes Beispiel für lokalisiertes, KI-gestütztes Infrastrukturmonitoring dar (weitere Informationen siehe Kasten 7).

Kasten 7. Costa Ricas *Contraloría General de la República* (CGR) dIAra – Überwachung von Infrastrukturprojekten

Das dIAra-Projekt stellt eine wegweisende Initiative dar, künstliche Intelligenz mit Internet-of-Things-(IoT)-Technologie zu integrieren, um die öffentliche Rechnungsprüfung von Infrastrukturprojekten zu verbessern. Das dIAra-Projekt wurde gemeinsam von der Universidad Cenfotec und dem CGR von Costa Rica mit Unterstützung der Deutschen Gesellschaft für Internationale Zusammenarbeit entwickelt und begegnet der Herausforderung, öffentliche Bauprojekte in Echtzeit zu überwachen. Es bietet Prüfer*innen automatisierte, datengestützte Erkenntnisse, die die Wirksamkeit und Transparenz der Aufsicht erhöhen.

Im Kern ist dIAra ein Überwachungssystem, das öffentliche Prüfer*innen unterstützt, indem es die Erfassung und Analyse von Bilddaten von Baustellen automatisiert. Kostengünstige IoT-Kameras, die vor Ort mit Mikrocontrollern und 3D-gedruckten Komponenten montiert werden, erfassen Bilder der laufenden Arbeiten. Diese Bilder werden von einem KI-gestützten Computer-Vision-Modell analysiert, das darauf trainiert ist, zentrale Elemente der Baustelle wie Arbeitskräfte und Ausrüstung zu erkennen. Die Ergebnisse der KI werden in Berichten strukturiert, die Anwesenheit und Aktivitätsniveaus auf der Baustelle ausweisen und es den Prüfer*innen ermöglichen, Arbeitsunterbrechungen oder Unregelmäßigkeiten zu erkennen, die eine weitere Untersuchung erfordern.

Das System ergänzt herkömmliche Prüfverfahren, die auf regelmäßigen Ortsbegehungen und manueller Berichterstattung basieren. Solche Ansätze können ressourcenintensiv sein und sind häufig im Umfang begrenzt, insbesondere bei groß angelegten oder geografisch verstreuten Projekten. Im Gegensatz dazu ermöglicht dIAra eine kontinuierliche und objektive Überwachung, ohne dass die physische Anwesenheit von Prüfer*innen erforderlich ist.

Aus technischer Sicht spiegelt die Entwicklung von dIAra das Bekenntnis zu technologischer Souveränität und offener Innovation wider. Das gesamte System, einschließlich Hardware-Designs, Software-Code, KI-Trainingsdatensätzen und Bereitstellungsdokumentation, ist Open Source. Diese Offenheit stellt sicher, dass die Plattform an lokale Bedürfnisse angepasst, von lokalen Technikern repariert werden kann und keine Abhängigkeit von Drittanbietern besteht. Ein solches Modell fördert nachhaltiges technologisches Wachstum in Costa Rica und kann von anderen Ländern oder Institutionen mit ähnlichen Aufsichtshürden problemlos übernommen werden.

Eine entscheidende Innovation in dIAra ist die KI-Komponente, insbesondere ein Modell namens *Modelo_1_29*, das ausschließlich mit Bildern von den überwachten Standorten trainiert wurde. Dieses Modell zeigte eine hohe Leistungsfähigkeit und erreichte einen durchschnittlichen Intersection-over-Union (IoU)-Wert von 82,7 Prozent sowie eine Sensitivitätsrate von 85,4 Prozent. Diese Kennzahlen weisen auf eine zuverlässige Erkennung von Objekten wie beispielsweise Arbeitskräften in Baustellenbildern unter den Bedingungen des Trainingsdatensatzes hin. Das KI-System erstellt automatisch strukturierte Ausgabedateien, die die erkannten Objekte zusammen mit Vertrauenswerten und räumlichen Koordinaten zusammenfassen. Dies ermöglicht eine effiziente Datenanalyse und unterstützt Prüfer*innen dabei, die tägliche Aktivität auf der Baustelle sowie Fortschrittmuster nachzuverfolgen.

Obwohl die Genauigkeit des Modells vielversprechend ist, erkennt das Projekt wichtige Einschränkungen an. Da die KI mit einem spezifischen Datensatz von einer einzelnen Baustelle trainiert wurde, kann ihre Leistungsfähigkeit nachlassen, wenn sie auf andere Baustellenumgebungen mit unterschiedlichen visuellen Merkmalen oder betrieblichen Bedingungen angewendet wird. Abweichungen bei der Beleuchtung, der Ausrüstung oder der Arbeitskleidung könnten die Genauigkeit des Modells beeinträchtigen. Nichtsdestotrotz erleichtert die Open-Source- und modulare Struktur des

Systems das erneute Training und die Anpassung, sodass zukünftige Implementierungen das Modell für unterschiedliche Kontexte weiter verfeinern können. Das Projekt befindet sich noch in der Entwicklung und erfordert eine weitere Klärung seiner Skalierbarkeit.

dIAra veranschaulicht, wie KI- und IoT-Technologien strategisch eingesetzt werden können, um Prüfungen öffentlicher Infrastrukturen zu modernisieren. Es bietet eine kosteneffiziente und replizierbare Lösung, die die Präzision und Frequenz der Überwachung verbessert, die institutionellen Aufsichtsfähigkeiten stärkt und zur technologischen Entwicklung auf nationaler Ebene beiträgt. Durch die Kombination automatisierter visueller Erkennung mit strukturierten Datenausgaben fördert dIAra ein Modell der transparenten, effizienten und skalierbaren Prüfung, das auch andere öffentliche Institutionen weltweit in Erwägung ziehen können, um eine verantwortungsvolle Verwaltung öffentlicher Ressourcen sicherzustellen.

Quelle: *Contraloría General de la República* (CGR) von Costa Rica, 2025

Der norwegische Riksrevisjonen untersuchte den Einsatz von Satellitenbildern in einer Wirtschaftlichkeitsprüfung mit Schwerpunkt auf Ernährungssicherheit. Ziel war es, zu beurteilen, wie viel Ackerland für die Nahrungsmittelproduktion genutzt wird. Mithilfe von Sentinel-2-Satellitendaten, die alle acht Tage aktualisiert werden, setzten Prüfer*innen Deep-Learning-Modelle auf Basis von CNNs ein, um die Kulturarten zu klassifizieren. Allerdings beeinträchtigte die anhaltende Bedeckung der Felder durch Wolken die Datenkonsistenz. Kommerzielle Satellitenbilder waren zwar verfügbar, wurden jedoch als unerschwinglich teuer eingestuft. Aufgrund dieser Einschränkungen wurden die Ergebnisse nicht vollständig in den abschließenden Prüfungsbericht aufgenommen, obwohl das Pilotprojekt das Potenzial KI-gestützter raumbezogener Analysen für die zukünftige Nutzung aufzeigte.

Wie diese Beispiele zeigen, bietet KI zwar neue Möglichkeiten, visuelle und räumliche Daten in die Prüfungsarbeit zu integrieren, ihr operativer Einsatz ist jedoch nach wie vor begrenzt. Die rechtlichen Rahmenbedingungen für den Einsatz von IoT-Geräten zur Überwachung unterscheiden sich erheblich zwischen den Rechtsordnungen, und ein verlässlicher Datenzugang bleibt eine zentrale Hürde. In vielen Rechtsordnungen setzen Datenschutz- und Privatsphäregesetze Grenzen dafür, wie IoT- und KI-Überwachungssysteme Informationen erfassen und weitergeben dürfen, insbesondere wenn daraus Rückschlüsse auf Aktivitäten von Einzelpersonen gezogen werden können. Anforderungen in Bezug auf Einwilligung, Transparenz und Datensicherheit schränken großflächige Überwachungen zusätzlich ein und können die Zulässigkeit oder Verlässlichkeit von sensorgestützten Beweismitteln in öffentlichen Prüfungen beeinflussen. Dennoch weisen diese Anwendungen auf ein zunehmendes Interesse hin, die Evidenzquellen für öffentliche Rechnungsprüfungsinstitutionen durch fortschrittliche Datenanalysetechniken zu erweitern. Weitere Anwendungen: interne Produktivitäts-, Kommunikations- und KI-Governance-Tools

Neben den zentralen Prüfungsfunktionen wird KI zunehmend eingesetzt, um die interne Produktivität von Prüfungsinstitutionen zu unterstützen, die Kommunikation zu erleichtern und Governance-bezogene Prozesse zu stärken. Auch wenn diese Anwendungen nicht unmittelbar zur Prüfungsarbeit beitragen, bieten sie praktische Vorteile in Bezug auf Effizienz, Wissensmanagement und institutionelle Agilität.

Stärkung der internen Kapazitäten und Produktivität

Mehrere Institutionen setzen KI-Tools ein, um interne Abläufe zu optimieren und die Produktivität der Mitarbeiter*innen zu steigern. Der EuRH prüft derzeit beispielsweise die Machbarkeit der Integration von KI-Komponenten in bestehende Robotic-Process-Automation (RPA)[†]-Lösungen, um Routineaufgaben im Zusammenhang mit der Prüfung zu unterstützen. Das britische NAO untersucht ebenfalls den Einsatz generativer KI-Tools wie Microsoft 365 Copilot, insbesondere für die Dokumentenprüfung, um Prüfer*innen

bei der effizienteren Bewältigung administrativer Aufgaben sowie bei weiteren Anwendungen zur persönlichen Produktivitätssteigerung zu unterstützen. Der Einsatz dieser Tools in den zentralen Prüfungsabläufen befindet sich noch in der Pilotphase.

Unterstützung der Kommunikation und öffentlichen Beteiligung

KI wird ebenfalls eingesetzt, um die institutionelle Kommunikation zu unterstützen, insbesondere bei der Erstellung von Inhalten für die Öffentlichkeit. Wie bereits zuvor erwähnt (siehe Generative KI für das Verfassen, Zusammenfassen und Übersetzen), hat der CGR in Costa Rica generative KI eingesetzt, um Drehbücher für Inhalte in sozialen Medien und Fernsehbriefings zu entwerfen. Diese Tools fördern eine flexible Content-Erstellung und werden vor der Veröffentlichung stets von Kommunikationsbeauftragten überprüft.

Beim Board of Audit and Inspection (BAI) in Korea wurden KI-gestützte Sentiment-Analyse-Tools pilotiert, um Medienberichte über deren Prüfungsergebnisse zu überwachen und so bisher manuelle Aufgaben zu automatisieren. Dies trägt dazu bei, nachzuvollziehen, wie Prüfungsthemen von der Öffentlichkeit aufgenommen und diskutiert werden, deckt verzerrte oder ungenaue Berichterstattung auf und stellt sicher, dass die Berichterstattung mit den ursprünglichen Pressemitteilungen der BAI übereinstimmt. Dieses Tool klassifiziert Medienberichte nach Tonalität (negativ, neutral oder positiv) und misst drei zentrale Aspekte:

- Ähnlichkeit: Wie genau Medienberichte die Pressemitteilungen widerspiegeln.
- Genauigkeit: Wie korrekt Probleme auf Grundlage der Pressemitteilungen berichtet werden.
- Stimmungswert: Der Gesamttenor der Medienberichte.

Etwa 16.000 Artikel aus verschiedenen Medien wurden analysiert. Die Ähnlichkeit wurde anhand der Kosinusähnlichkeit gemessen, die Genauigkeit mittels Named Entity Recognition (NER) bewertet und die Stimmungswerte klassifizierten die Berichte in negative, neutrale oder positive Kategorien (weitere Informationen siehe Kasten 8).

KI für Steuerung, Aufsicht und Politik-analyse

KI entwickelt sich ebenfalls als Tool für Governance und Politikanalyse. Beim Europäischen Rechnungshof wurde KI im Rahmen einer Prüfungsaufgabe getestet, um Gesetzesdokumente zu vergleichen. Die menschliche Überwachung und Überprüfung der Ergebnisse bleibt entscheidend, doch eröffnet dieser Ansatz neue Möglichkeiten zur Analyse des Gesetzgebungsprozesses und zum Verständnis seiner institutionellen Auswirkungen.

Kasten 8. Board of Audit and Inspection (BAI) in Korea: Eine KI-basierte Methodik zur Analyse von Ergebnissen interner Prüfungen

Das BAI von Korea entwickelt „Leitlinien zur Erstellung interner Prüfungspläne“, um die Effizienz der Prüfungen und die Zusammenarbeit zwischen Behörden zu verbessern. Derzeit sind etwa 200.000 Prüfungsakten im Public Audit Portal gespeichert, doch die Analyse erfolgt weiterhin manuell, was die Skalierbarkeit und Effizienz einschränkt. Um diesem Umstand zu begegnen, hat das BAI ein Projekt gestartet, um interne Prüfungsfälle mithilfe von KI zu klassifizieren und die Selbstprüfungspraktiken in den Kommunalverwaltungen zu bewerten. Das Projekt konzentriert sich auf die Prüfungsergebnisse 2024 und umfasst vier Phasen: 1) Entwicklung von Klassifizierungsstandards, 2) Datenerhebung und -vorverarbeitung, 3) Fallanalyse und 4) Aufbau eines KI-basierten Klassifizierungsmodells. Das Audit- und Inspektionsforschungsinstitut und das Amt für proaktive Verwaltungsdienste sowie die interne Revision des öffentlichen Sektors führten das Projekt in Zusammenarbeit durch.

1. Entwicklung von Klassifizierungsstandards

Ein Klassifizierungsrahmen wurde entwickelt, um die interne Kategorisierung von Prüfungsfällen bei BAI zu standardisieren, wobei anerkannt wurde, dass die Einstufung je nach Perspektive der einzelnen Forscher*innen variieren kann. Der Rahmen berücksichtigte Rückmeldungen von provinziellen Prüfungsabteilungen und Praktikern vor Ort, um Objektivität und Vollständigkeit sicherzustellen.

2. Datenerhebung

Aufgrund von Zugriffsbeschränkungen wurde ein spezieller Crawler entwickelt, um 24.954 öffentlich verfügbare Prüfberichte (HWP, PDF) zu erfassen. Diese wurden für eine strukturierte Analyse verarbeitet.

3. Fallanalyse

Rund 1.200 Berichte wurden manuell geprüft, um zentrale Merkmale zu extrahieren und Trainingsdaten zu erstellen. Argument Mining (eine NLP-Technik, die argumentative Strukturen aus Texten extrahiert), wurde eingesetzt, um aus unstrukturierten Texten logische Schlussfolgerungen zu gewinnen. Die Daten wurden bereinigt (Duplikate entfernt, Fehler korrigiert usw.) und vor der Verwendung in ein einheitliches Format standardisiert.

4. Entwicklung von KI-Modellen

Ein Ensemble-Modell, das Entscheidungsbaum- und Naive-Bayes-Algorithmen zur Steigerung der Vorhersagegenauigkeit kombiniert, wurde darauf trainiert, Prüfungsfälle zu klassifizieren. Das Modell übertraf die Einzelsysteme, indem es komplementäre Stärken nutzte: Entscheidungsbäume sind insbesondere besser darin, komplexe Variablenbeziehungen zu bewältigen, während Naive-Bayes-Modelle probabilistische Vorhersagen ermöglichen. Zusätzlich wurde das Llama-Sprachmodell eingesetzt, um den Nicht-Trainingsdatensatz automatisch zu kennzeichnen und zusammenzufassen. Das Projekt hat zum Ziel, das ML-Ensemble-Modell mit dem LLM zu vergleichen, um festzustellen, welche Methode die besten Ergebnisse liefert; die Validierung ist derzeit im Gange.

Ein Prototyp wurde entwickelt, um die praktische Anwendung des Modells zu demonstrieren und die Machbarkeit zu überprüfen. Der Prototyp wurde so konzipiert, dass er die Interessen tatsächlicher BAI-Nutzer*innen widerspiegelt, indem er Vergleiche des internen Prüfungsstatus zwischen verschiedenen Regionen ermöglicht und detailliertere Analysen einzelner Prüfungsfälle bereitstellt.

Quelle: Board of Audit and Inspection (BAI) in Korea, 2025.

Diese Anwendungsfälle zeigen eine große Vielfalt potenzieller Einsatzmöglichkeiten von KI in Prüfungsinstitutionen, sei es durch den Einsatz von ML, neuronalen Netzen, NLP oder generativer KI. Obwohl sich die meisten Initiativen noch im Pilot- oder Erkundungsstadium befinden, zeigt die Vielfalt der Ideen und Experimente, dass die Organisationen das transformative Potenzial dieser Ansätze erkennen.

Bei allen Anwendungsfällen stehen menschliche Prüfer*innen weiterhin im Mittelpunkt des Prozesses – von der Konzeption über die Erprobung bis hin zur Einführung. Diese Betonung spiegelt sowohl praktische Erwägungen wider (um Akzeptanz und Vertrauen in den frühen Einsatzphasen zu gewährleisten) als auch eine bewusste Reaktion auf aktuelle Herausforderungen im Zusammenhang mit algorithmischer Transparenz und Nachvollziehbarkeit, die in einem späteren Abschnitt erörtert werden. KI wird daher als Tool eingesetzt, um das fachliche Urteilsvermögen zu unterstützen und nicht zu ersetzen.

Die Konsultationen zeigten zudem, dass die Wirkung von KI verstärkt werden kann, wenn sie mit ergänzenden Technologien wie benutzerfreundlichen Schnittstellen, Datenvisualisierung, robotergestützter Prozessautomatisierung und IoT-Anwendungen kombiniert wird. Eine solche Integration ermöglicht eine effizientere Datenverarbeitung, umfassendere analytische Erkenntnisse und Benutzerfreundlichkeit sowie optimierte Prüfungsabläufe.

In diesem Stadium ist der schnellste Fortschritt bei der Nutzung von KI für prüfungsunterstützende Funktionen wie Dokumentenanalyse oder Risikostichproben zu beobachten, nicht jedoch bei der Generierung unmittelbarer Prüfungsergebnisse. Im Vergleich zu anderen Aufsichts- oder Prüfungsbehörden setzen SAIs KI tendenziell seltener für ex-ante Risikobewertungen oder groß angelegte Datenanalysen ein. Dieser Unterschied spiegelt zum Teil ihre ex-post-Mandate und die ungleiche Verfügbarkeit digitaler Daten in den geprüften Einrichtungen wider.

Insgesamt deuten die Ergebnisse darauf hin, dass sich der Einsatz von KI in der Prüfung zwar noch im Anfangsstadium befindet, die Grundlagen für eine breitere Anwendung jedoch aktiv ausgelotet werden. Im nächsten Abschnitt werden die Infrastruktur, die Kompetenzen und die institutionellen Rahmenbedingungen untersucht, die zur Unterstützung dieser wachsenden KI-Ambitionen entwickelt werden.

4 Infrastruktur und Kapazität

In diesem Kapitel wird eine institutionsübergreifende Analyse der digitalen und KI-Fähigkeiten der befragten öffentlichen Rechnungshöfe vorgestellt. Die Bewertung stützt sich ausschließlich auf selbstberichtete Daten und konzentriert sich auf vier zentrale Dimensionen: Hardware-Infrastruktur, Software-Umgebungen, Datenmanagement und -governance sowie Entwicklungskapazitäten. Ziel ist es, zu verstehen, wie Prüfungsinstitutionen sich darauf vorbereiten und positionieren, KI zur Unterstützung ihrer Kernaufgaben einzusetzen, indem gemeinsame Ansätze und Abweichungen identifiziert werden.

Hardware-Infrastruktur

Die Hardware-Infrastruktur der befragten Institutionen variiert erheblich hinsichtlich ihres Reifegrads, ihrer Skalierbarkeit und ihrer Ausrichtung auf die Ziele des KI-Einsatzes. Ein wiederkehrendes Thema ist das Bemühen, technologische Experimente mit strengen Datenschutzanforderungen in Einklang zu bringen. Während einige Organisationen stark in lokale digitale Infrastrukturen investiert haben (z. B. lokale Datenspeicherung und Rechenleistung), haben sich andere für cloudbasierte Dienste von kommerziellen Anbietern oder hybride Modelle entschieden, um den sich entwickelnden Anforderungen an künstliche Intelligenz gerecht zu werden.

Mehrere Institutionen haben sich dafür entschieden, die Infrastruktur intern aufzubauen und zu betreiben. Spezielle Rechenumgebungen, darunter leistungsfähige Workstations und GPU-Server, wurden angeschafft, um KI-Experimente und -Entwicklungen zu unterstützen. Diese Konfigurationen umfassen häufig isolierte Umgebungen mit internen Hochgeschwindigkeitsnetzwerken, die sicheres Experimentieren mit ML-Modellen ermöglichen und dabei eine kontrollierte Umgebung gewährleisten. In einigen Fällen wurden dafür eigens physische Räume vorgesehen. Diese Räume sind so eingerichtet, dass Internetzugang erlaubt ist, jedoch der Zugang zum internen Netzwerk eingeschränkt bleibt, was eine absichtliche, risikobewusste Maßnahme darstellt, um Innovation zu fördern, ohne die Systemsicherheit zu gefährden.

Tatsächlich wird in den Institutionen die Verfügbarkeit von Hardware für Experimente als ein Motor für interne Innovation angesehen. Teams werden häufig ermutigt, neue KI-Ideen in sicheren Umgebungen zu erproben, wobei einige Zugang zu Tools wie LM Studio oder PrivateGPT für das Ausführen lokaler großer Sprachmodelle bereitstellen. Diese Tools bieten eine „Sandbox“-Umgebung, die Daten-Teams und Prüfer*innen die Möglichkeit gibt, Funktionen zu testen, ohne das Risiko von Datenlecks oder einer Abhängigkeit von kommerziellen APIs einzugehen.

Einige Organisationen beginnen, in spezialisierte Hardware zu investieren, um fortgeschrittene KI-Workloads zu unterstützen. Dazu gehört der Aufbau von GPU-Clustern für das interne Training oder die Feinabstimmung von Sprachmodellen. In einigen Fällen werden sowohl cloudbasierte als auch lokale Lösungen gleichzeitig erprobt, um die effizienteste langfristige Konfiguration zu finden.

Einige der befragten Organisationen setzen Hochleistungs-GPUs wie NVIDIA H200 oder L40 ein, was auf einen Wandel hin zu fortschrittlicheren KI-Anwendungen über grundlegende Automatisierungen oder Reporting/Dashboarding hinaus hinweist. Während einige Institutionen eigene Hardware erwerben, bevorzugen andere Institutionen, GPUs zu mieten (d. h. GPU-Kapazitäten über die Cloud zu nutzen) und

greifen dabei auf kommerzielle Cloud-Anbieter zurück. Dieser Ansatz bietet Flexibilität und Skalierbarkeit, insbesondere für Institutionen, die ihre KI-Strategien noch entwickeln oder vorübergehend Rechenkapazität benötigen. Das Mieten ermöglicht es Organisationen zudem, verschiedene Modellarchitekturen zu erproben, ohne sich auf bestimmte Hardware-Investitionen festlegen zu müssen. Dieses Vorgehen kann jedoch zu Kostenschwankungen und betrieblichen Abhängigkeiten führen, die sorgfältig gesteuert werden müssen.

Eine Mehrheit der Organisationen hat Bedenken hinsichtlich der Datensouveränität und der Rechtsvorschriften für das Hosting und die Verarbeitung von Daten geäußert, die ihre Vorgehensweisen maßgeblich beeinflussen. Einige verzichten ausdrücklich auf globale kommerzielle Cloud-Dienste aufgrund geopolitischer, datenschutzrechtlicher oder rechtlicher Erwägungen und bevorzugen lokale Rechenzentren oder regional verwaltete Plattformen. Selbst unter denjenigen, die große internationale Cloud-Anbieter nutzen, besteht häufig die Präferenz für Infrastrukturen, die sich innerhalb der eigenen nationalen oder kontinentalen Grenzen befinden.

Ergänzend zu diesen Entwicklungen greifen SAls, die sich in den frühen Phasen der KI-Einführung befinden, in der Regel auf gemeinsame staatliche Infrastrukturen oder leichtgewichtige virtuelle Umgebungen zurück, um erste Anwendungsfälle zu testen und Prototypen zu entwickeln. Ein bemerkenswertes Beispiel ist ein speziell für Prüfer*innen entwickeltes sicheres virtuelles Desktop-System, das den Fernzugriff auf abgeschlossene Datenanalysetools ermöglicht. Dies spiegelt ein übergeordnetes Muster wider, bei dem Hardware-Entscheidungen nicht nur von Leistungsanforderungen, sondern auch von Compliance- und Risikomanagement-Überlegungen bestimmt werden.

Nur eine geringe Zahl von Institutionen berichtete von klaren Plänen, Maßnahmen zu ergreifen, die eine Erweiterung oder Modernisierung ihrer Hardware-Infrastruktur in naher Zukunft erfordern würden, wie etwa die Implementierung von Spark-Clustern (eine Art Cluster-Computing-Plattform) oder den Aufbau von Data Lakes zur Unterstützung intensiverer Rechenvorgänge. Dies deutet darauf hin, dass die KI-Experimentierung zwar an Fahrt gewinnt, erhebliche Investitionen in die Hardware-Infrastruktur jedoch weiterhin auf wenige fortschrittliche Institutionen beschränkt sind.

Im Allgemeinen verfügen diejenigen, die in ihrer KI-Entwicklung weiter fortgeschritten sind, über ausgereifere Infrastrukturen. Diese Organisationen verfügen nicht nur über die physische Hardware, um anspruchsvolle Modelle zu betreiben, sondern haben auch Verfahren entwickelt, die einen sicheren, effizienten und regelkonformen Betrieb im Einklang mit bestehenden regulatorischen Rahmenbedingungen ermöglichen. Umgekehrt kann es bei Organisationen, die stark auf externe Hosting-Dienste oder allgemeine Infrastrukturen angewiesen sind, in manchen Fällen zu Einschränkungen kommen, wenn die Anwendungsfälle für KI komplexer und datenintensiver werden. Dennoch konnten Organisationen, die einen strategischen Ansatz bei der Nutzung kommerzieller Tools verfolgen, erfolgreich KI-Suiten aufbauen, ohne zusätzliche lokale Infrastruktur bereitstellen zu müssen, indem sie über kommerzielle Partner cloudbasierte Zugänge zu Tools, Rechenleistung und Datenspeicherung nutzen.

Es gilt als bewährte Praxis, sicherzustellen, dass die Infrastrukturentscheidungen auf die spezifischen Anforderungen der aktuellen und angestrebten KI-Workloads abgestimmt sind, die in der Regel sowohl eine hohe Verfügbarkeit von Rechenleistung als auch robuste Maßnahmen zur Datensicherheit erfordern. Institutionen, die ihre KI-Fähigkeiten ausbauen möchten, können davon profitieren, hybride Architekturen zu entwickeln, die die Flexibilität der Cloud mit der Kontrolle und Compliance von On-Premises-Systemen verbinden. Dies würde sowohl Experimente als auch den produktionsreifen Einsatz ermöglichen und dabei Kosten, Sicherheit und operative Agilität in Einklang bringen.

Insgesamt haben zwar mehrere SAls eine solide Grundlage für eine KI-fähige Infrastruktur geschaffen, dennoch besteht weiterhin eine erhebliche Lücke zwischen ersten Experimenten und skalierbaren, produktionsreifen Umgebungen. Künftige Investitionen werden voraussichtlich darauf abzielen müssen, diese Lücke zu schließen, insbesondere wenn Prüfungsinstitutionen von Pilotprojekten mit KI-Tools zur Integration in zentrale Prozesse übergehen.

Softwareumgebung und -stack

Eine Vielzahl unterschiedlicher Softwaretools und Plattformen wird in den Institutionen eingesetzt, um die digitale Transformation und die Integration von KI zu unterstützen. Die Software-Ökosysteme unterscheiden sich in ihrer Tiefe und Spezialisierung, was häufig den Reifegrad der übergeordneten Digitalstrategie der jeweiligen Organisation widerspiegelt. Insgesamt zeichnet sich ein deutlicher Trend ab, kommerzielle Produktivitätssuiten und Open-Source-Frameworks zu nutzen und diese gezielt durch die Erprobung spezialisierterer KI-Softwareumgebungen zu ergänzen.

Nahezu alle Institutionen setzen auf etablierte kommerzielle Plattformen für Produktivität und Zusammenarbeit, wobei Microsoft 365 und Google Workspace am häufigsten genutzt werden. Diese Tools werden häufig mit integrierten KI-Assistenten wie Microsoft 365 Copilot angeboten, die zunehmend in größerem Umfang getestet oder eingeführt werden. Institutionen, die diese Funktionen integriert haben, berichten von Verbesserungen der Arbeitsabläufe, obwohl einige Einschränkungen eingeführt wurden, um die Nutzung auf öffentliche oder nicht sensible Daten zu beschränken. Diese Schutzmaßnahmen sind üblich, um Bedenken hinsichtlich Datenlecks auszuräumen, insbesondere wenn Inhalte extern verarbeitet oder für das Training von Modellen Dritter verwendet werden könnten.

Open-Source-Frameworks wie Python, Jupyter Notebooks und zugehörige Data-Science-Bibliotheken werden breit eingesetzt, insbesondere in Teams, die sich mit praktischer Analytik und Automatisierung beschäftigen. Mehrere Prüfungsinstitutionen unterhalten GitHub-Repositorien für Versionskontrolle, interne Zusammenarbeit und Projektverfolgung. Diese Praktiken weisen auf einen Reifegrad in der Softwareentwicklung hin, der iterative Experimente und Wiederverwendbarkeit von Code unterstützt. In einigen Fällen werden Codebasen intern als strategische Vermögenswerte aufgebaut und gepflegt, wodurch die Grundlage für gemeinsam genutzte KI-Komponenten in verschiedenen Prüfungsfunktionen geschaffen wird.

Fortgeschrittenere Organisationen haben begonnen, KI-spezifische Tools und Plattformen wie LM Studio, PrivateGPT oder RAG-Architekturen einzusetzen, um LLMs in kontrollierten Umgebungen auszuführen – entweder zu Testzwecken oder für den eingeschränkten Produktionseinsatz – und dabei ein höheres Maß an Kontrolle über Datenschutz und Leistungsfähigkeit zu gewährleisten. Einige Organisationen haben zudem untersucht, kleinere Sprachmodelle gezielt auf prüfungsbezogene Terminologie und regionale Sprachen abzustimmen oder zu trainieren. Diese Initiativen spiegeln ein wachsendes Bewusstsein für den Wert domainspezifischer KI-Tools im Vergleich zu allgemeinen Modellen wider.

Cloud-native Dienste sind ebenfalls Teil der Softwarelandschaft. Einige Institutionen, die komplexere Datenverarbeitung und Modellbereitstellung anstreben, experimentieren mit AWS Bedrock oder integrieren mit Palantir Foundry. Die Einführung kommerzieller KI-Plattformen erfolgt jedoch weiterhin mit Zurückhaltung und beschränkt sich häufig auf Pilotprogramme oder isolierte Implementierungen. Institutionen bleiben weiterhin sensibel gegenüber Fragen der Datenresidenz und des Vendor-Lock-ins, und viele bevorzugen Softwarelösungen, die Datenhoheit und operative Transparenz gewährleisten.

Sicherheit und Compliance bleiben entscheidende Faktoren bei der Auswahl von Software. Institutionen haben Richtlinien eingeführt, um sicherzustellen, dass keine personenbezogenen Daten extern übermittelt oder zum erneuten Training von Modellen Dritter (wie Microsoft 365 Copilot) verwendet werden. Einige haben lokale Richtlinien oder die Aufsicht durch den Chief Information Security Officer (CISO) eingeführt, um zu genehmigen, welche Tools verwendet werden dürfen, insbesondere wenn diese mit sensiblen Datensätzen interagieren. Die Nutzung von LLMs über interaktive Tools wie Online-Notebooks (z. B. Jupyter, Databricks) oder sichere, in sich geschlossene Umgebungen (z. B. Docker), die im Webbrowser ausgeführt werden, verdeutlicht, wie Organisationen die Benutzerfreundlichkeit mit dem Erfordernis der Kontrolle über Daten und Systeme in Einklang bringen.

Dateninfrastruktur und Governance

Eine entscheidende Grundlage für jede erfolgreiche KI- oder digitale Transformationsinitiative ist die Qualität, Zugänglichkeit und Governance von Daten. Unter den befragten Institutionen variieren die Reifegrade erheblich: Einige haben stark in leistungsfähige Data Lakes und strukturierte Datenpipelines investiert, während andere sich noch in den Anfängen der Angleichung unterschiedlicher Systeme und der Etablierung von Standards für die Datenverwaltung befinden, wie es in der (2019^[4]) Arbeit der OECD zur Datenverwaltung im öffentlichen Sektor hervorgehoben wird.

Mehrere Organisationen haben zentrale Data Lakes aufgebaut, die Informationen aus verschiedenen internen und externen Quellen konsolidieren, darunter Finanzdaten, Unternehmensregister, Daten zu öffentlichen Vergaben und andere hochwertige Datensätze. In einer Institution ist der Data Lake auf über 44 Terabyte angewachsen – also groß genug, um Millionen von Dokumenten zu speichern – und unterstützt Hunderte von Nutzer*innen sowie nahezu sämtliche Prüfungstätigkeiten. Diese weiterentwickelten Umgebungen werden durch Metadatenmanagementsysteme, automatisierte Risikoindikatoren und leistungsstarke Data-Mining-Funktionen ergänzt. Solche Institutionen haben häufig Jahre, manchmal über ein Jahrzehnt, darauf verwendet, die technischen und organisatorischen Grundlagen für dieses Maß an Datenintegration zu schaffen.

Es gibt zudem zahlreiche Beispiele von Institutionen, die Systeme entwickeln, um strukturierte und unstrukturierte Daten miteinander zu verbinden. Wenn Dokumente in Formaten wie PDF-Dateien oder gescannten Bildern eingehen, werden Python-Skripte und individuell entwickelte Extract-Transform-Load-(ETL)-Routinen[†] eingesetzt, um die Inhalte zu extrahieren und für die Analyse zu strukturieren. In einem Fall verwenden Prüfteams eigens entwickelte Werkzeuge, um Rohdateien in strukturierte Tabellen umzuwandeln, sodass Prüfer*innen mit unterschiedlichen Datensätzen arbeiten können, ohne auf komplexe Analysewerkzeuge angewiesen zu sein. Dies spiegelt einen praxisorientierten, anwendungsfallbasierten Ansatz für Dateninfrastrukturen wider, bei dem Effizienz und Prüfbereitschaft Vorrang vor der Abhängigkeit von einer allgemeinen Datenarchitektur haben.

Auch die Richtlinien zur Daten-Governance unterscheiden sich. Mehrere Institutionen verfügen über formelle Datenstrategien, die häufig intern oder in Zusammenarbeit mit externen Berater*innen entwickelt werden. Diese Strategien legen in der Regel besonderen Wert auf Datenqualität, Standardisierung und sicheren Zugang. Einige Institutionen haben bereits begonnen, Governance-Ebenen rund um ihre KI-Tools zu entwickeln, um sicherzustellen, dass die menschliche Aufsicht über algorithmische Ergebnisse gewahrt bleibt.

Eine wiederkehrende Herausforderung in mehreren Organisationen ist die Fragmentierung der Datenquellen. Altsysteme, isolierte Datensätze und uneinheitliche Formate sind häufige Probleme. So steht beispielsweise eine Institution vor dem Problem überlappender Datenbanken und nicht aufeinander abgestimmter Strukturen, insbesondere bei Verträgen und öffentlichen Aufzeichnungen. Dies verdeutlicht die Schwierigkeiten bei der horizontalen Integration und weist darauf hin, dass selbst Datensätze aus demselben Ministerium hinsichtlich Umfang und Struktur voneinander abweichen können. Diese Komplexität wird dadurch verstärkt, dass SAls häufig auf Daten angewiesen sind, die nicht in ihrem Besitz sind (z. B. Daten von geprüften Stellen), sodass sie deren Qualität nicht kontrollieren können. Darüber hinaus unterliegen Prüfungsthemen häufigen Veränderungen, was zu erheblichen Änderungen der für die Analyse benötigten Datensätze führt. Infolgedessen erfolgen Datenintegration und -bereinigung häufig nachträglich und erfordern erheblichen manuellen Aufwand. Diese Herausforderungen behindern die Automatisierung und schränken die Skalierbarkeit jeglicher Art von Analysen ein. Als Reaktion darauf investieren einige Institutionen/Regierungen in Initiativen zur Standardisierung und Zentralisierung. Ein Beispiel betrifft ein gemeinsames Prüfungsmanagementsystem, das derzeit entwickelt wird, um Arbeitsabläufe in allen Behörden zu harmonisieren und die Konsistenz der eingehenden Daten zu verbessern. Eine weitere Institution testet ein System, das Dashboards und Analyse-Skripte in einer

einzigsten Umgebung integriert und so reibungslosere Übergänge von der Rohdatenerfassung bis zum Endergebnis ermöglicht. Diese Initiativen verdeutlichen ein zunehmendes Verständnis dafür, dass die digitale Transformation ebenso sehr – wenn nicht sogar mehr – von einer konsistenten Dateninfrastruktur abhängt wie von neuartigen KI-Anwendungen.

Es werden ebenfalls Anstrengungen unternommen, um sicherzustellen, dass die Datenqualität während des gesamten Lebenszyklus erhalten bleibt. Dies umfasst nicht nur die technischen Aspekte der Datenaufnahme und -transformation, sondern auch organisatorische Prozesse, die die Verantwortung für die Datenintegrität zuweisen. In einigen Institutionen wurde die Position eines Chief Data Officer oder einer vergleichbaren Funktion eingeführt, um eine zentralisierte Datenpolitik voranzutreiben und die Datenkatalogisierung zu überwachen. Dies spiegelt einen umfassenderen institutionellen Wandel wider, bei dem Daten als strategischer Vermögenswert betrachtet werden.

Während einige Institutionen noch dabei sind, Teams oder Richtlinien für die Datenverwaltung aufzubauen, experimentieren andere bereits mit fortschrittlicheren Datenarchitekturen. Dazu gehören Vektordatenbanken[†] für den Einsatz mit großen Sprachmodellen sowie die Integration mit Diensten wie OpenSearch und semantischen Indexierungsprotokollen. Solche Entwicklungen sind nach wie vor relativ selten, deuten jedoch auf einen Übergang zu stärker KI-nativen Dateninfrastrukturen hin.

Im Allgemeinen scheinen die SAIs, die frühzeitig einen Schwerpunkt auf Daten-Governance gelegt haben, bei der Einführung von KI schneller voranzukommen. Ihre strukturierten, zugänglichen und gut dokumentierten Datensätze bieten eine solide Grundlage für ML und Automatisierung. Im Gegensatz dazu stellen Institutionen mit fragmentierten oder nicht dokumentierten Datenquellen häufig fest, dass selbst grundlegende Analysen einen unverhältnismäßigen Aufwand erfordern, wodurch das Potenzial für skalierbare KI-Tools eingeschränkt wird.

Globale Standards verlangen zunehmend von öffentlichen Institutionen, Datenumgebungen bereitzustellen, die sicher, interoperabel und prüfbar sind. Da sich KI weiterentwickelt, werden starke Daten-Governance und Infrastruktur weiterhin die wichtigsten Treiber digitaler KI-basierter Innovation bleiben. Institutionen, die auf langfristige Interoperabilität und Governance statt nur auf kurzfristige Anwendungsfälle setzen, werden in den kommenden Jahren am ehesten in der Lage sein, anspruchsvollere Lösungen zu übernehmen. Da Rechnungsprüfungsbehörden zunehmend auf KI setzen, um Erkenntnisse und Kontrollen zu steuern, wird die Integrität der Daten, auf denen diese Tools basieren, noch entscheidender werden.

Entwicklungskapazitäten

Die Fähigkeit von Institutionen, KI- und digitale Lösungen effektiv zu entwickeln und einzusetzen, hängt eng mit ihren internen Kompetenzen, Organisationsstrukturen und dem Zugang zu spezialisiertem Fachwissen zusammen. In der befragten Gruppe gibt es erhebliche Unterschiede in der Organisation der Entwicklungsarbeit, die von vollständig internen Teams bis hin zu hybriden Modellen reichen, die auf externe Berater*innen oder Forschung und Partnerschaften zurückgreifen. Diese Unterschiede spiegeln verschiedenen Reifegrade im Bereich KI, Budgetbeschränkungen und strategische Prioritäten wider.

Mehrere Institutionen haben sich für die Eigenentwicklung als ihr primäres Modell entschieden, was häufig durch Bedenken hinsichtlich Datensicherheit, Kontrolle des geistigen Eigentums und langfristiger Nachhaltigkeit motiviert ist. Diese Institutionen unterhalten in der Regel spezialisierte Teams für Data Science oder Innovation, die damit beauftragt sind, KI-Tools zu entwickeln, zu testen und bereitzustellen, die direkt auf die Anforderungen der Rechnungsprüfung abgestimmt sind. In solchen Fällen arbeiten Entwicklungsteams eng mit Prüfer*innen zusammen, um deren Herausforderungen und Ziele zu identifizieren und zu verstehen sowie zu beurteilen, ob ein KI-Anwendungsfall oder eine andere Herangehensweise eine praktikable und optimale Lösung darstellt. Diese Teams bestehen häufig aus

einer Mischung von Data Scientists, Entwickler*innen und Fachexpert*innen, die mitunter von internen Innovationslaboren oder Kompetenzzentren unterstützt werden, die als institutionelle Knotenpunkte für Experimente dienen.

Bei den fortschrittlicheren Organisationen hat die interne Entwicklung zur Schaffung maßgeschneiderter KI-Tools geführt, die speziell auf Rechnungsprüfungsaufgaben zugeschnitten sind. Beispiele umfassen Systeme zur Risikopriorisierung, Erkennung von Preisabweichungen, Dokumentenzusammenfassung, Vertragsprüfung und sogar RAG-Chatbots zur Unterstützung bei der Auslegung von Prüfungsstandards. Diese Lösungen werden häufig mithilfe von Open-Source-Frameworks oder internen Plattformen entwickelt und ermöglichen ein hohes Maß an Anpassung. Institutionen, die diesen Ansatz verfolgen, betonen die Bedeutung der Integration von Entwicklungszyklen in den übergeordneten organisatorischen Arbeitsablauf, anstatt digitale Tools als eigenständige technische Projekte zu behandeln.

Im Gegensatz dazu arbeiten mehrere Institutionen mit hybriden Modellen, die internes Fachwissen mit externer Unterstützung kombinieren, in der Regel durch die Beschaffung von Entwicklungs- und Beratungsleistungen oder durch Partnerschaften mit Unternehmen oder anderen Forschungseinrichtungen. Dies entspricht im Allgemeinen den Ansätzen, die in Regierungsorganisationen insgesamt verfolgt werden (OECD, 2025^[1]). Solche Vereinbarungen entstehen häufig, wenn interne Teams klein oder noch im Aufbau sind und externe Partnerschaften den Zugang zu spezialisiertem Fachwissen oder eine beschleunigte Projektdurchführung ermöglichen. In solchen Fällen konzentriert sich das interne Personal in der Regel auf Projektüberwachung, Integration und langfristige Wartung, während externe Berater*innen die anfängliche Entwicklung oder Skalierung übernehmen. Zusammenarbeit mit akademischen Einrichtungen, KI-Laboren oder großen Beratungsunternehmen ist üblich, insbesondere wenn Institutionen ihre Kompetenzen langfristig aufbauen oder komplexe technische Anforderungen bewältigen möchten.

In Institutionen, in denen keine eigenen KI-Teams existieren, wird die Entwicklung häufig von engagierten Einzelpersonen oder kleinen Gruppen technisch versierter Mitarbeiter*innen vorangetrieben, die oftmals nach einem Bottom-up-Innovationsmodell arbeiten. In solchen Umgebungen erkennen Mitarbeiter*innen – häufig jüngere Beschäftigte oder kürzlich eingestellte externe Fachkräfte – Automatisierungs- oder Effizienzpotenziale und schlagen digitale Tools zu deren Umsetzung vor. Die Unterstützung durch das Management für solche Initiativen variiert, hat jedoch in einigen Fällen zur Entstehung informeller Innovationscluster geführt, die sich schließlich zu formalisierten Rollen oder Abteilungen entwickeln. In diesen Kontexten ist die Unterstützung durch das Management entscheidend, sowohl um eine Organisationskultur zu schaffen, die Innovation wertschätzt, als auch um aktiv die Identifizierung, Validierung und Umsetzung hochwertiger Ideen zu ermöglichen. Dadurch kann von unten angestoßene Innovation über bloße Ad-hoc-Experimente hinausgehen (OECD, 2025^[1]).

Dennoch kann das Fehlen strukturierter Entwicklungsteams erhebliche Herausforderungen mit sich bringen. Institutionen ohne eigene Fachkompetenz könnten Schwierigkeiten haben, über Pilotprojekte hinauszugehen, insbesondere wenn ihnen die internen Ressourcen fehlen, um extern entwickelte Lösungen zu betreiben oder zu skalieren. In solchen Kontexten besteht selbst für vielversprechende Proof-of-Concepts das Risiko, aufgrund sich wandelnder Softwareumgebungen, nicht gepflegter Codebasen oder Personalfuktuation obsolet zu werden. Darüber hinaus kann die Abhängigkeit von externen Partnern ohne klare Pläne für den Wissenstransfer oder die Sicherstellung der Eigentumsrechte an geistigem Eigentum und Daten die Fähigkeit der Institution beeinträchtigen, ihre Tools an sich ändernde Prüfungsprioritäten anzupassen oder weiterzuentwickeln.

Einige Institutionen haben versucht, diesem Umstand zu begegnen, indem sie Entwicklungskompetenzen in bestehende Abteilungen wie IT, Datenanalyse oder Innovationsbüros integriert haben. In einem Fall wird erwogen, zwei separate Teams – eines mit Schwerpunkt auf Softwareentwicklung und Datenengineering, das andere auf Analytik – zusammenzulegen, um die Koordination zu verbessern und Doppelarbeit zu vermeiden. Andere SAls befinden sich noch in einem früheren Stadium und arbeiten mit

externen Forschungsinstituten zusammen, um grundlegende Fähigkeiten zu entwickeln. In diesen Fällen besteht das langfristige Ziel in der Regel darin, internes Fachwissen aufzubauen, selbst wenn die anfänglichen Fortschritte stark von externen Partnerschaften abhängen.

Bei den untersuchten Institutionen sind jene KI-Initiativen am erfolgreichsten, bei denen technisches Fachwissen, Domainkenntnisse und Unterstützung durch die Führungsebene eng miteinander verzahnt sind. Entwicklungskompetenz ist nicht nur eine Frage der personellen Ressourcen oder der technischen Infrastruktur, sondern beruht auf institutioneller Abstimmung, Governance und einem gemeinsamen Bekenntnis zur Innovation.

Die Analyse zeigt ein vielfältiges Bild der digitalen Reife unter den teilnehmenden Institutionen. Während einige Organisationen erhebliche Investitionen in zentrale Dateninfrastrukturen, sichere Entwicklungsumgebungen und interne KI-Kompetenzen getätigt haben, befinden sich andere noch in der Aufbauphase grundlegender Elemente oder sind auf externe Partnerschaften angewiesen, um ihre Kapazitäten auszubauen. Zu den häufigsten Herausforderungen zählen fragmentierte Datensysteme, begrenzte interne Fachressourcen und der Bedarf an klareren Governance-Strukturen für den Einsatz von KI. Gleichzeitig gibt es deutliche Anzeichen für Innovation, Zusammenarbeit und institutionelles Engagement bei der Integration von KI in die Funktionen der öffentlichen Rechnungsprüfung. Für die Zukunft wird es entscheidend sein, weiterhin in interne Entwicklungsteams, eine solide Daten-Governance und sektorübergreifendes Lernen zu investieren, um sicherzustellen, dass KI-Lösungen sowohl wirkungsvoll als auch mit den Werten des öffentlichen Sektors vereinbar sind.

5 Wirkungsmessung

Die Messung der Auswirkungen bei der Einführung von KI in der öffentlichen Rechnungsprüfung entwickelt sich zu einer zentralen Herausforderung und steht im Einklang mit den allgemeinen Entwicklungen im Bereich KI in der öffentlichen Verwaltung (OECD, 2025^[1]). Obwohl das Interesse daran wächst, den Return on Investment von KI-Tools zu verstehen, ist der Zusammenhang zwischen deren Einsatz und greifbaren Ergebnissen häufig indirekt, zeitlich verzögert oder schwer isolierbar. In vielen Fällen ermöglichen KI-Tools, insbesondere generative KI-Modelle, neue Arten von Aufgaben, die zuvor nicht praktikabel waren, wie etwa das systematische Durchsuchen großer Dokumentenmengen oder das Erstellen von Inhalten außerhalb der Arbeitszeiten von Prüfer*innen. Diese Formen der Produktivitätssteigerung sind wertvoll, aber nicht leicht quantifizierbar. Ebenso können Verbesserungen bei der Prüfungsqualität oder -tiefe tatsächlich vorhanden sein, lassen sich jedoch mit den derzeitigen Messgrößen nur schwer erfassen.

Diese Schwierigkeit wird zudem dadurch verstärkt, dass der Einsatz von KI und die damit verbundene Qualifizierung in den meisten SAls noch in den Anfängen steckt. Infolgedessen haben die Institutionen bislang noch keine umfassenden Rahmenwerke entwickelt, um die Auswirkungen dieser Tools zu bewerten. In vielen Fällen laufen Experimente und Pilotprojekte weiter, und gewonnene Erkenntnisse werden schrittweise integriert. In Ermangelung belastbarer Kennzahlen könnten SAls erwägen, strukturierte qualitative Rückmeldungen (z. B. Nutzerzufriedenheitsumfragen, Erfahrungsberichte von Prüfer*innen) einzubeziehen, um den wahrgenommenen Nutzen und die Benutzerfreundlichkeit zu erfassen. Ein vergleichendes Benchmarking mit gleichrangigen Institutionen oder früheren Prüfungszyklen könnte ebenfalls dazu beitragen, informelle, aber umsetzbare Indikatoren für Fortschritte zu schaffen.

Dennoch beginnen einige Organisationen damit, praktischere Indikatoren für Fortschritt und Mehrwert zu definieren. Eine Institution beispielsweise verfolgt Kennzahlen wie:

- den Prozentsatz der Abteilungen der Institution, die KI-Tools aktiv nutzen,
- die Häufigkeit der Nutzung von KI-Tools,
- Time-to-Value (die Zeitspanne bis zur Ergebniserzielung nach der Implementierung),
- die Anzahl der geschulten Mitarbeiter*innen und/oder
- die Ablösung von Altsystemen durch KI-gestützte Arbeitsabläufe.

Dies stellt weiterhin eine erste Ebene der Bewertung der weitverbreiteten Nutzung von KI in den Kernfunktionen dar; jedoch handelt es sich dabei noch nicht zwangsläufig um eine fortgeschrittene Analyse des ROI oder der Auswirkungen des Einsatzes von KI.

Frühe Fortschritte wurden auch im Hinblick auf die Zeitersparnis bei Wissensmanagement und Schreibunterstützung festgestellt. In den befragten Institutionen, in denen LLMs auf standardisierte Prüfungsdokumentation trainiert werden, hat der Einsatz generativer KI die schnelle Erstellung von Erstentwürfen von Berichten ermöglicht, was im Vergleich zu herkömmlichen Schreibmethoden zu erheblichen Zeitersparnissen geführt hat. Die Möglichkeit, Materialien außerhalb der regulären Arbeitszeiten zu verarbeiten – beispielsweise durch die Automatisierung von Prozessen über Nacht – wurde ebenfalls als bedeutender Produktivitätsgewinn hervorgehoben.

Es sind auch ehrgeizigere Ziele entstanden. Eine Organisation berichtete ein klares Ziel, ihren Prüfungsstellenbestand um 50 Stellen zu reduzieren und dabei KI und Automatisierung zur Umsetzung

dieses Wandels einzusetzen. Solche Ziele sind zwar selten, verdeutlichen jedoch, dass einige Institutionen den Einsatz von KI nicht nur zur schrittweisen Verbesserung, sondern für grundlegende Veränderungen in den Prüfungsabläufen prüfen.

Auch außerhalb des öffentlichen Sektors bieten Erfahrungen privater Rechnungsprüfungsgesellschaften wertvolle Einblicke. Ein Unternehmen hob hervor, dass durch die Kombination von KI-Tools und Prozessstandardisierung die Prüfungsdurchlaufzeiten für geprüfte Stellen von 60 Tagen auf nur noch 10 bis 15 Tage gesenkt werden konnten. In diesem Fall ist jedoch nicht allein die Technologie ausschlaggebend: Erst das Zusammenspiel von Tools, die Standardisierung von Daten und die Neugestaltung von Arbeitsabläufen sowie die Weiterbildung des Personals ermöglichen diese Fortschritte. Dies untermauert die Erkenntnis, dass tatsächliche Wirkung aus der Abstimmung von Tools, Prozessen und Menschen entsteht und nicht aus dem isolierten Einsatz von Technologie.

Um Regierungen dabei zu unterstützen, Herausforderungen in diesem Bereich besser zu bewältigen, entwickelt die OECD einen Leitfaden zur Messung von Auswirkungen und zum Nachweis des Nutzens von KI-Investitionen (OECD, forthcoming^[5]).

6 Herausforderungen und strategische Überlegungen

In diesem Kapitel werden die zentralen Herausforderungen und strategischen Ausrichtungen für die Einführung von KI in der öffentlichen Rechnungsprüfung dargestellt, wie sie aus den Konsultationen hervorgegangen sind. Institutionen äußerten durchgehend Bedenken hinsichtlich Ethik, Nachvollziehbarkeit und Modellverzerrung. Viele nannten auch die organisatorische Trägheit als hemmenden Faktor für Innovationen. Datenverwaltung und -sicherheit stellten ebenfalls eine bedeutende Herausforderung dar, die durch die Notwendigkeit verstärkt wurde, mit der schnellen technologischen Entwicklung Schritt zu halten und die Abhängigkeit von externen Anbietern zu steuern. Zusammen weisen diese Perspektiven auf strategische Ansätze hin, die sowohl bei der technischen Umsetzung als auch in den organisatorischen Rahmenbedingungen auf Flexibilität und Anpassungsfähigkeit ausgerichtet sind.

Ethik, Erklärbarkeit und Verzerrung

In den Konsultationen äußerten Organisationen durchgehend Bedenken hinsichtlich Ethik, Integrität und Transparenz, insbesondere im Zusammenhang mit KI-gestützten Tools. Die Wahrung des öffentlichen Vertrauens erfordert, dass KI, die in Prüfungen eingesetzt wird, nicht nur effektiv, sondern auch vertretbar, transparent und nachvollziehbar ist. Mehrere Organisationen wiesen auf Risiken wie Halluzinationen, algorithmische Voreingenommenheit (Bias) und mangelnde Erklärbarkeit von KI-Ergebnissen hin; obwohl diese Risiken unterschiedlicher Art sind, könnten sie alle die Glaubwürdigkeit und Validität von Prüfungsergebnissen gefährden.

So wurde beispielsweise in Gesprächen mit der Niederländischen *Algemene Rekenkamer* hervorgehoben, dass aktuelle LLMs unterschiedliche Sprachen und Kulturen besser abbilden könnten, die über jene hinausgehen, die den großen Anbietern entsprechen. Sprachbarrieren verschärfen diese Herausforderungen und schränken die Anwendbarkeit solcher Modelle in nicht-englischsprachigen Kontexten ein. Um diesem Umstand zu begegnen, wartet die niederländische Algemene Rekenkamer auf die Entwicklung eines niederländischsprachigen Modells, das eine höhere sprachliche und kulturelle Relevanz bieten soll, um so Fairness und Vertrauenswürdigkeit bei KI-gestützten Prüfungen zu stärken. Ein solches Beispiel ist GPT-NL, das rund 13,5 Mio. EUR investiert, die vom niederländischen Ministerium für Wirtschaft und Klima (EZK) bereitgestellt werden, um ein Modell zu trainieren (OECD, 2025^[11]).

Um den damit verbundenen Herausforderungen entgegenzuwirken, konzentrieren sich viele Organisationen derzeit darauf, KI in Bereichen einzusetzen, die Prüfer*innen unterstützen und die Produktivität steigern, in denen – wie auch bei Menschen – eine gewisse Fehlermarge akzeptabel ist. Beispielsweise kann KI dabei helfen, große Mengen an Dokumentation effizienter zu prüfen als es Menschen allein möglich wäre, ohne dass die Prüfungsergebnisse direkt von den KI-Ausgaben abhängen. Die letztendliche Verantwortung verbleibt weiterhin bei den Prüfer*innen und stellt zugleich eine gängige Maßnahme zur Risikominderung dar, um sicherzustellen, dass alle Ergebnisse geprüft und mit dem erforderlichen Maß an fachlicher Urteilsfähigkeit und Skepsis behandelt werden. Wie ein Interviewpartner anmerkte: „Halluzinationen sind ein Feature, kein Bug“; da es derzeit unmöglich ist, sie vollständig

auszuschließen, müssen Prüfer*innen darin geschult werden, KI-generierte Ergebnisse zu erkennen, kritisch zu bewerten und zu validieren.

Mehrere Institutionen hoben das Fehlen von Prüfungsstandards für den Umgang mit von KI erzeugten oder analysierten Beweismitteln hervor. Dies spiegelt breitere Trends wider, bei denen Regierungsorganisationen keine umsetzbaren Leitlinien dafür haben, wie KI in bestimmten Funktionen und Bereichen vertrauenswürdig eingesetzt werden kann (OECD, 2025^[1]). In der Praxis bedeutet dies, dass KI-generierte Erkenntnisse derzeit der gleichen kritischen Prüfung unterliegen wie solche, die mit traditionellen Methoden gewonnen wurden, jedoch ohne klare Leitlinien zur Validierung oder Dokumentation ihrer Zuverlässigkeit. Das Fehlen von Standards führt zu Unsicherheit – sowohl bei den Prüfer*innen und den Qualitätssicherungsstellen als auch in der Öffentlichkeit.

Als Reaktion auf Fragen zur Transparenz der Ergebnisse hat der norwegische Riksrevisjonen zudem eine Forschungspartnerschaft mit einer Universität initiiert, um die Erklärbarkeit von KI weiterzuentwickeln. Die Erklärbarkeit von Algorithmen wird heute ebenfalls dadurch behandelt, dass mit Modellen gearbeitet wird, die eine gewisse Transparenz bieten, indem sie Berichte generieren, die ihre Ergebnisse und deren Quellen erläutern (weitere Informationen siehe Kapitel 7. Gute Praktiken und Erfolgsfaktoren).

Offene Datenpraktiken, bei denen Daten und Softwarecode der Öffentlichkeit frei zugänglich gemacht werden, tragen dazu bei, Transparenz und Rechenschaftspflicht sicherzustellen. Durch die Veröffentlichung anonymisierter Datensätze und Codes ermöglichen Organisationen die unabhängige Überprüfung von Ergebnissen und fördern damit das Vertrauen in den Prozess. Die Einführung offener Datenstandards, die Bereitstellung klarer Dokumentationen zu Datenquellen und Methodologien sowie die Einrichtung unabhängiger Überprüfungsmechanismen können Transparenz und Vertrauen weiter fördern. Die KI-Prinzipien der OECD (2024^[6]), die von 48 Ländern und der Europäischen Union anerkannt wurden, bieten wertbasierte Grundsätze und Empfehlungen für politische Entscheidungsträger*innen für den vertrauenswürdigen Einsatz von KI, einschließlich der in diesem Abschnitt behandelten Themen. Der OECD-Rahmen für vertrauenswürdige KI in der öffentlichen Verwaltung (Abbildung 2) bietet einen Ansatz zur Anwendung dieser Prinzipien im öffentlichen Sektor.

Organisationskultur, Kompetenzen und Unterstützung der Einführung

Die Einführung von KI befindet sich bei vielen befragten öffentlichen Rechnungskontrollbehörden weiterhin in einem frühen Stadium. Mehrere Organisationen berichteten von Widerstand gegen Veränderungen und einer langsamen Anpassung an KI- und datenbasierte Ansätze im Allgemeinen, insbesondere unter Prüfer*innen, die nicht ausreichend im effektiven Einsatz von KI-Tools geschult sind.

Eine effektive Schulung der Nutzer*innen ist für die erfolgreiche Einführung von KI-Tools in der Prüfung unerlässlich. Unzureichende Schulungen können zu unsachgemäßer Nutzung sowie zu Ängsten und Risikoscheu führen, was die Einführung behindert. Das Thema Prompt-Design (d. h. wie Anweisungen in klarer Sprache formuliert werden, um eine Reaktion eines KI-Systems zu erhalten) wurde in mehreren Gesprächen angesprochen, insbesondere im Zusammenhang mit der Herausforderung, Nutzer*innen darin zu schulen, effektive Anweisungen für Chatbots und große Sprachmodelle zu erstellen. Tatsächlich verringert eine schlechte Gestaltung von Anweisungen nicht nur die Wirksamkeit eines Tools, sondern kann auch zu Frustration bei den Nutzer*innen sowie zu einer langsameren oder geringeren Akzeptanz führen. Um dem entgegenzuwirken, entwickeln SAIs und andere Prüfungsorganisationen Schulungsprogramme für ihre Mitarbeiter*innen (teilweise verpflichtend) und bieten kontinuierliche Weiterbildung an. Die Schulungen für Nutzer*innen konzentrieren sich derzeit hauptsächlich auf den Einsatz generativer KI, während Weiterbildungsmaßnahmen in anderen Bereichen der KI überwiegend auf IT- und Datenexpert*innen ausgerichtet sind.

Die Bedeutung von Führung und Governance wurde als unterstützend für Veränderungsprozesse hervorgehoben, was laut einer OECD-Untersuchung (2025^[1]) als entscheidend für eine erfolgreiche Einführung von KI in jeder Regierungsorganisation angesehen wird. Insbesondere das britische NAO beschrieb einen aktiven Ansatz zur KI-Governance durch:

- Einrichtung eines KI-Governance Board zur Identifizierung, Bewertung und Steuerung von Risiken, die sich aus neuen KI-Technologien und Anwendungsfällen ergeben
- Einstellung eines neuen Direktors für KI und Datenanalyse
- Einführung von Bewertungs- und Überprüfungsprozessen im Zusammenhang mit dem Einsatz neuer Technologien wie KI innerhalb des Qualitätsmanagementsystems
- Bereitstellung von Schulungs- und Leitfäden für Praktiker zum verantwortungsvollen Einsatz von KI.

Der norwegische Riksrevisjonen rechnet mit Herausforderungen im Bereich des kulturellen Wandels und sieht Führungskompetenz als entscheidend an, um organisatorische Veränderungen voranzutreiben und die Weiterbildung der Prüfer*innen zu unterstützen. Sie hoben außerdem einen Ansatz für das Veränderungsmanagement hervor, bei dem Change Agents eingesetzt werden, um einen internen Kulturwandel hin zu einer stärker datenbasierten Arbeitsweise zu fördern.

Datenqualität als strategischer Wert

Obwohl die zunehmende Verfügbarkeit von Daten den Einsatz von Technologie ermöglicht, bleibt die Datenqualität eine grundlegende Herausforderung bei der Einführung von KI. Prüfungsspezifische Aufgaben erfordern strukturierte, saubere und häufig standardisierte Daten, doch viele SAs haben mit fragmentierten, inkonsistenten oder unstrukturierten Datensätzen zu kämpfen, insbesondere in prüfungsrelevanten Bereichen mit hohem Textanteil. Uneinheitliche Datenformate zwischen geprüften Stellen, Jahren, Abteilungen und Institutionen können in KI-gestützten Prüfungen zu Fehlern führen, da Datenbeschränkungen die Genauigkeit und Präzision der Modelle beeinträchtigen.

Ein privates Prüfungsunternehmen stellte fest, dass die Harmonisierung von Daten eine Priorität darstellt, wobei Pilotmodelle für die Datenbereinigung und die Einrichtung von Arbeitsabläufen entwickelt werden. Für SAs gestaltet sich die Herausforderung noch komplexer, da sie in der Regel nicht Eigentümer der von ihnen geprüften Daten sind. Stattdessen müssen sie sich auf die Datenmanagementpraktiken anderer Regierungsstellen verlassen oder nachträgliche Tools einsetzen, um die Daten für die Analyse aufzubereiten. Um diese Herausforderungen zu bewältigen, investieren Organisationen in Tools zur Datenvorverarbeitung und -bereinigung, etablieren Daten-Governance-Rahmenwerke und fördern, wo möglich, die Standardisierung von Daten, um deren Qualität und Nutzbarkeit zu verbessern. Um Engpässe zu vermeiden, ist die Übereinstimmung zwischen den in das Organisationssystem eingehenden Daten und den Prüfungsmanagement-Tools oder -Systemen von entscheidender Bedeutung, insbesondere bei uneinheitlicher Datenverarbeitung wie Zeitstempeln.

Governance und Sicherheit

Sicherheit ist ein zentrales Anliegen bei der Einführung von KI in der Rechnungsprüfung. Da leistungsstarke kommerzielle KI-Produkte zunehmend verfügbar und erschwinglich werden, haben viele SAs die Nutzung cloudbasierter Lösungen als besondere Herausforderung identifiziert, insbesondere im Hinblick auf Datensicherheit sowie die Einhaltung gesetzlicher und regulatorischer Rahmenbedingungen.

Sicherheit und Compliance stehen auch im Mittelpunkt der Daten-Governance-Bemühungen, wobei mehrere Organisationen unter strengen Vorgaben arbeiten, die die Nutzung personenbezogener oder

sensibler Daten in KI-Tools einschränken. In einigen Fällen führt die Einhaltung von Vorschriften und Richtlinien zur lokalen Bereitstellung von KI-Modellen, um die mit der Übertragung von Daten über Landesgrenzen hinweg oder in Drittsysteme verbundenen rechtlichen Risiken zu vermeiden.

Beispielsweise hat der BRH in Deutschland ein striktes Prinzip der Dateneigentümerschaft eingeführt. Dateneigentümer können bis zu einem bestimmten Vertraulichkeitsgrad eigenständig über die Nutzung von Daten in externen Cloud-basierten Tools wie Microsoft Copilot entscheiden. Andere haben sichere interne Umgebungen implementiert, wie beispielsweise isolierte virtuelle Desktop-Umgebungen, um die Kontrolle über den Datenzugriff und den Datenstandort (den physischen oder geografischen Speicherort der Daten) zu gewährleisten. Bei einigen befragten Institutionen haben Datenschutz- und geopolitische Bedenken zu einer umfassenderen Neubewertung von Cloud-Lösungen geführt. Beispielsweise wird beim norwegischen Riksrevisjonen ein Wechsel hin zu lokalen Infrastrukturen vorangetrieben, um die Abhängigkeit von ausländischen Cloud-Dienstleistern zu verringern.

Dennoch ist die Nutzung kommerzieller Plattformen weiterhin verbreitet. Einige nutzen umfassende vertragliche Vereinbarungen mit großen Anbietern wie Microsoft, um den Datenschutz und die Datenintegrität zu gewährleisten. In allen Institutionen gilt die Umsetzung robuster Datensicherheitsmaßnahmen sowie die Einhaltung nationaler oder regionaler Datenschutzvorschriften und -gesetze, sofern zutreffend, als unerlässlich, um Datensicherheit und Compliance bei der Entwicklung von KI zu gewährleisten.

Management des Veränderungstempos und Reduzierung der Abhängigkeiten von Drittanbietern

Die rasante Entwicklung von KI-Tools stellt für die SAls ein strategisches Dilemma dar. Technologien, die heute eingeführt werden, können schon bald veraltet sein oder nicht mehr unterstützt werden, was Fragen hinsichtlich Altsystemen, langfristiger Wartung und Nachhaltigkeit sowie kontinuierlicher Schulungsanforderungen für das Personal aufwirft. Eine Institution beschrieb diese Dynamik als ein „Wettrüsten“ zwischen öffentlichen Prüfungsstellen und Technologieanbietern, bei dem die Aufrechterhaltung relevanter Kompetenzen ständige Investitionen und Wachsamkeit erfordert.

Das Tempo des Wandels birgt nicht nur finanzielle Risiken, sondern auch strategische und operative Konsequenzen. Häufige Wechsel oder Aktualisierungen von Tools können zu Fragmentierung, Doppelarbeit und dem, was einige als „Tool-Wildwuchs“ bezeichneten, führen – eine Situation, in der zu viele voneinander unabhängige Tools eher Verwirrung und Ineffizienzen als Klarheit schaffen.

Mehrere SAls äußerten Bedenken hinsichtlich einer zu starken Abhängigkeit von kommerziellen Plattformen. Kommerzielle Plattformen ermöglichen es Prüfungsinstitutionen, mit dem technologischen Fortschritt Schritt zu halten, erleichtern den Zugang zu Fachwissen und technischen Lösungen und vereinfachen die Wartung. Allerdings können kurze Produktlebenszyklen, wechselnde Prioritäten der Anbieter und Herausforderungen bei der Datensicherheit dazu führen, dass SAls anfällig für Störungen sind oder an einen bestimmten Anbieter gebunden werden.

Um diesem Problem zu begegnen, streben einige SAls, wie beispielsweise in Korea, an, auf ein Modell der internen Entwicklung umzusteigen, um die Abhängigkeit von Drittanbietern zu verringern und die Kontrolle über zentrale Systeme zurückzugewinnen. Dies erfordert jedoch erhebliche Rechenressourcen, technisches Fachpersonal sowie Wartungskapazitäten, über die viele SAls derzeit nicht verfügen. Das Feintuning von LLMs, beispielsweise zur Anpassung an auditspezifische Aufgaben, ist derzeit stark durch die verfügbare Rechenleistung und Fachkompetenz eingeschränkt.

Ausgewogenheit zwischen Innovation, langfristiger Stabilität und Nachhaltigkeit

Vor dem Hintergrund dieser Herausforderungen suchen Institutionen wie das Accountant General's Department (AGD) in Singapur nach Entwicklungsmodellen, die ein Gleichgewicht zwischen Experimentierfreude und operativer Stabilität herstellen und dabei die Kosten der digitalen Transformation berücksichtigen.

Die Sicherstellung der langfristigen Tragfähigkeit und Skalierbarkeit von KI-Lösungen bleibt eine zentrale Herausforderung, insbesondere die Fähigkeit, Systeme während des gesamten Lebenszyklus eines KI-Systems intern zu betreiben und weiterzuentwickeln. Diese Herausforderung kann durch modulare Technologien und Designs mit austauschbaren Komponenten abgemildert werden, da sie schnellere Upgrades, reibungslosere Aktualisierungen und eine effizientere Erweiterung ermöglichen. Dadurch wird die Flexibilität erhöht und Ausfallzeiten werden minimiert. Die Einbettung strukturierter Feedbackschleifen in die Entwicklungs-, Nutzungs- und Optimierungszyklen von Tools stärkt die Anpassungsfähigkeit zusätzlich und stellt sicher, dass die Tools sich kontinuierlich an die Bedürfnisse der Nutzer*innen sowie an veränderte Rahmenbedingungen anpassen. Ergänzend dazu ist ein kontinuierlicher interner Kapazitätsaufbau unerlässlich, um Institutionen mit den technischen und organisatorischen Fähigkeiten auszustatten, die erforderlich sind, um KI-Lösungen langfristig zu steuern und anzupassen. Ohne anpassungsfähiges Design und kontinuierliches Feedback besteht selbst für technisch ausgereifte Tools das Risiko, in realen Prüfungsabläufen an Relevanz zu verlieren.

Um diese langfristigen Stabilitäts- und Nachhaltigkeitsbemühungen zu unterstützen, ist es entscheidend, neue Tools nach Möglichkeit integriert und ergänzend mit bereits genutzter Software-Infrastruktur einzubinden oder gut durchdachte Veränderungsmanagement-Maßnahmen zu ergreifen, um bestehende Ansätze zu ersetzen. Dies trägt dazu bei, eine nahtlose Integration in bestehende Arbeitsabläufe oder deren Transformation sicherzustellen und gleichzeitig den Wert früherer Investitionen zu maximieren. Ein Beispiel hierfür findet sich beim britischen NAO, wo ein Proof of Concept erfolgreich LLMs auf ein bestehendes Tool aufgesetzt hat, um im Archiv veröffentlichter Prüfungen zu recherchieren (siehe Kasten 9). Der Erfolg dieser Integrationen hängt maßgeblich vom Anpassungsgrad der Systeme und der Verfahrensflexibilität ab.

Kasten 9. Das britische National Audit Office (NAO): Integration von LLMs für semantische Suche und Antwortgenerierung

Beim britischen NAO befindet sich die Integration von LLMs in eine bestehende interne Anwendung, den „Back Catalogue Analyser“, derzeit in der Pilotphase. Mit dieser Erweiterung sollen semantische Suchfunktionen und KI-gestützte, generative Antworten in natürlicher Sprache in ein Tool integriert werden, das bislang auf herkömmlichen, schlagwortbasierten Suchmethoden basiert.

Die bestehende Anwendung dient als Suchtool für das Backkatalog-Archiv, um frühere Veröffentlichungen zu finden und zu analysieren. Obwohl das Tool in seiner bisherigen Form effektiv war, beschränkt es sich darauf, nach exakten Wörtern oder Wortgruppen innerhalb der Berichtstexte zu suchen. Das geplante Upgrade stellt eine bedeutende Verbesserung dar, da es eine intuitivere und intelligendere Informationssuche ermöglicht.

Durch die Einbindung von LLMs wird das Tool in der Lage sein, die semantische Bedeutung hinter den Nutzeranfragen zu erfassen und dadurch präzisere sowie kontextbezogene Suchergebnisse zu ermöglichen. Diese Erweiterung wird zudem die Generierung von Antworten in natürlicher Sprache ermöglichen, sodass Nutzer*innen mit dem System in einem Dialog interagieren können. Dadurch können Nutzer*innen relevante Informationen effizienter abrufen, ohne Anfragen umformulieren oder sich um die exakte Formulierung sorgen zu müssen.

Die Integration von LLMs wird voraussichtlich mehrere zentrale Vorteile bieten, darunter eine Verringerung der übersehenen relevanten Inhalte aufgrund von Synonymen oder unterschiedlichen Formulierungen, eine verbesserte Verarbeitung komplexer Suchausdrücke sowie die Automatisierung von Aufgaben wie der Zusammenfassung und Themenextraktion aus Berichtstexten. Dies wird nicht nur den Rechercheprozess optimieren, sondern auch den derzeit erforderlichen manuellen Aufwand für Aufgaben wie Zusammenfassung und Themenextraktion erheblich verringern.

Quelle: UK National Audit Office, 2025.

7 Gute Praktiken und Erfolgsfaktoren

Während Prüfungsinstitutionen weltweit die Integration von KI in Prüfverfahren untersuchen, wurden im Rahmen der Konsultationen verschiedene Ansätze deutlich, die Organisationen, die vergleichbare Wege erwägen, wertvolle Erkenntnisse bieten könnten. Die befragten Institutionen haben zentrale Erfolgsfaktoren und bewährte Verfahren geteilt. Diese Erkenntnisse, gewonnen aus unterschiedlichen institutionellen Erfahrungen, verdeutlichen potenzielle organisatorische, technische und kulturelle Strategien zur Nutzung von KI, wobei die Prüfungsintegrität und das Vertrauen der Interessengruppen besonders betont werden.

Transparenz und Nachvollziehbarkeit

Wie zuvor erörtert, haben sich Transparenz und Nachvollziehbarkeit von KI-Ergebnissen als grundlegende Elemente bei der Einführung von KI-Tools herauskristallisiert und stellen tatsächlich eines der fünf wertebasierten Prinzipien der OECD-KI-Prinzipien dar (2024^[6]). Die SAls betonten die Bedeutung transparenter KI-Praktiken und erklärbarer KI-Ergebnisse, um Vertrauen zu wahren und zugleich die Akzeptanz durch Prüfer*innen zu erleichtern. Die Aufrechterhaltung eines Prüfprotokolls und die Nachvollziehbarkeit der Entscheidungsfindung hinter KI-Ergebnissen sind unerlässlich; Institutionen wiesen wiederholt darauf hin, dass aktuelle Modelle häufig nicht vollständig erklärbar sind. Eine Institution betonte, dass bei der Auswahl zwischen verschiedenen Modellen stets die Erklärbarkeit gegenüber der Leistungsfähigkeit Vorrang hat.

Einige Institutionen begegnen Herausforderungen in Bezug auf Erklärbarkeit durch die Organisation ihrer Arbeitsabläufe. In erster Linie, wie bereits festgestellt, führen die meisten dargestellten Anwendungsfälle nicht eigenständig zu Prüfungsergebnissen, sondern dienen vielmehr der Effizienzsteigerung.

Bei der brasilianischen CGU wurde das Chatbot-Modell so entwickelt, dass es den zugrunde liegenden Bericht anzeigt, der zur Generierung der Antwort verwendet wurde, um so die Nachverfolgbarkeit zu verbessern. In ähnlicher Weise prüft der EuRH den Einsatz von Retrieval-Augmented Generation (RAG)-Architekturen, um sicherzustellen, dass KI-Ergebnisse ausschließlich auf der eigenen Dokumentation und den eigenen Daten der Institution basieren und so die Verbindung zwischen Eingaben und KI-generierten Inhalten stärken.

Wesentlich ist, dass sich die Schulung der Nutzer*innen bislang als wirksamste Schutzmaßnahme erwiesen hat, um den sachgemäßen Einsatz von KI-Tools sicherzustellen. Schulungsprogramme konzentrieren sich auf den angemessenen Einsatz von KI, betonen die fortwährende professionelle Skepsis gegenüber KI-generierten Ergebnissen und unterstreichen, dass KI nicht als alleinige Grundlage für Prüfungsfeststellungen dienen sollte. Viele Institutionen setzen sich dafür ein, dass das Personal aufmerksam bleibt, indem KI-Ergebnisse kontinuierlich mit den Originalquellen abgeglichen werden. Als bewährte Praxis in einigen Institutionen sind Prüfer*innen zudem verpflichtet, systematisch offenzulegen, dass KI zur Generierung von Ergebnissen eingesetzt wurde, um die Transparenz weiter zu fördern.

Einbindung von Nutzer*innen und Interessengruppen: Förderung von Innovationen von unten nach oben

In den Konsultationen wurden verschiedene Ansätze für die Entwicklung und Einführung von KI festgestellt. Einige Institutionen verfolgen eine Bottom-up-Strategie, die von engagierten Nutzer*innen vorangetrieben wird; andere orientieren sich an der strategischen Ausrichtung der Führungsebene und des oberen Managements, während viele beide Ansätze miteinander kombinieren. In allen Fällen scheint eine ausgeprägte Kultur der kontinuierlichen Einbindung von Nutzer*innen und Interessenträger*innen entscheidend für den Erfolg zu sein.

Der Grad der Einbindung der Nutzer*innen, die nahezu immer Prüfer*innen sind, in den Entwicklungsprozess der Tools variiert in der Praxis. In einigen Institutionen spielen Prüfer*innen eine zentrale Rolle bei der Identifizierung operativer Schwachstellen und der Definition technischer Anforderungen. Dieses partizipative Modell stellt sicher, dass digitale Werkzeuge weiterhin auf die tatsächlichen Bedürfnisse ausgerichtet sind und von den Mitarbeiter*innen an vorderster Front leichter übernommen werden. In anderen Fällen beginnen die Prüfer*innen selbst, Skripte zu schreiben oder bestehenden Code anzupassen, insbesondere in Organisationen, in denen digitale Kompetenzen weiter verbreitet sind oder in denen kürzlich graduierte Prüfer*innen mit ausgeprägten digitalen oder Programmierkenntnissen in den Mitarbeiterstab eintreten. Diese Demokratisierung von KI-Entwicklungsaktivitäten spiegelt die zunehmende Zugänglichkeit von Data-Science-Tools sowie die wachsende Erwartung wider, dass Fachkräfte unterschiedlicher Disziplinen direkt mit digitalen Arbeitsabläufen arbeiten.

Bottom-up-Innovation war ein entscheidender Treiber für die Einführung von KI in Institutionen wie der CGR Costa Ricas und dem EuRH. In Costa Rica, wo die Institution noch relativ jung ist, entwickelte sich ein organischer Ansatz für Innovation. Mitarbeiter*innen aus verschiedenen Funktionen, die Motivation und vielversprechende Ideen zeigten, erhielten bezahlten Zugang zu einem kommerziellen KI-Tool, um dieses zu erkunden und anzuwenden. Die Wahl einer kostenpflichtigen Standardplattform förderte die Nutzung, da sie Bedenken hinsichtlich Datenschutz und Sicherheit ausräumte, die häufig mit kostenlosen Tools verbunden sind (weitere Informationen siehe Kasten 10).

Auch beim EuRH konzentrierte sich die erste Einführung von KI darauf, engagierte Mitarbeiter*innen aus verschiedenen Funktionen zu befähigen. Frühe KI-Anwender*innen waren mit ihrer Begeisterung und Vielfalt an Perspektiven maßgeblich daran beteiligt, einen Bottom-up-Prozess der Ideenfindung und des Experimentierens zu fördern.

Im Kontext partizipativer Innovation haben sich agile Methoden als förderlich für die Einbindung von Nutzer*innen in digitale Projekte erwiesen. Beispielsweise verfolgt die niederländische Algemene Rekenkamer einen sehr agilen Ansatz bei der Entwicklung von Tools: Es werden Prototypen erstellt, diese den Prüfer*innen beziehungsweise Nutzer*innen zum Testen zur Verfügung gestellt, Feedback eingeholt und die Tools in einem iterativen Zyklus weiterentwickelt, der die Entwicklung digitaler Lösungen fördert.

Das AGD von Singapur verfolgt einen schrittweisen Ansatz zum Aufbau von Kompetenzen im Bereich der Digitalisierung und des Einsatzes von KI-Technologien für die interne Revision. Die internen Rechnungsprüfer*innen werden ermutigt, Pilotprojekte oder Anwendungsfälle zu initiieren, um potenzielle Vorteile zu erkunden. Wenn es sinnvoll ist, erfolgreiche Anwendungsfälle auszuweiten, geschieht dies durch den Austausch bewährter Verfahren und Tools innerhalb des Berufsstands.

Die Ernennung von Produktverantwortlichen mit vorheriger Prüfungserfahrung zur Leitung der Entwicklung von KI-Pilotprojekten und -Tools wurde als weiterer Erfolgsfaktor identifiziert. Diese hybriden Profile (Rechnungsprüfung + Digital) tragen dazu bei, die Lücke zwischen der technischen Entwicklung von KI und den praktischen Anforderungen der Prüfungsfunktionen zu überbrücken. Wenn solche hybriden Profile fehlen, kann Bottom-up-Innovation dennoch wirksam durch einen pragmatischen Ansatz gefördert werden:

indem Personen mit spezifischen Bedürfnissen mit jenen vernetzt werden, die Lösungen entwickeln oder bereitstellen können. Offene Feedback-Mechanismen mit schnellen Reaktionszyklen und einer „No-Blame“-Kultur, kombiniert mit realistischen Erwartungen und kontinuierlicher Unterstützung, fördern eine kontinuierliche Verbesserung und stärken einen effektiven Bottom-up-Innovationsprozess.

Unterstützung durch Management und Führungsebene

Starke Führung und aktive Unterstützung durch das Management sind entscheidende Faktoren für eine erfolgreiche Einführung von KI, insbesondere in Situationen, in denen Innovation durch Risikoscheu oder bürokratische Trägheit gehemmt wird. Führung trägt dazu bei, den Rahmen für Experimente zu setzen, Ressourcen bereitzustellen und den Einsatz neuer Technologien sowie die digitale Erkundung zu legitimieren. Die Unterstützung von oben gewährleistet darüber hinaus, dass KI-Projekte mit den institutionellen Zielen in Einklang stehen.

Im Fall von Costa Rica beispielsweise hat das obere Management KI-Projekte aktiv unterstützt und so ein förderliches Umfeld geschaffen, das Kreativität, Experimentierfreude und kalkulierte Risikobereitschaft begünstigt (siehe Kasten 10).

Das Ausmaß des institutionellen Ehrgeizes prägt ebenfalls die Entwicklungspfade von KI. Während einige Organisationen einen vorsichtigen, schrittweisen Ansatz verfolgen – mit dem Schwerpunkt auf der Ausweitung erfolgreicher Pilotprojekte oder dem Aufbau von Infrastrukturen für Experimente –, haben andere ehrgeizigere Ziele gesetzt. Eine Institution hat beispielsweise das Ziel gesetzt, den Personalbestand in der Prüfung durch den Einsatz von KI zu reduzieren, was ein entschlossenes Bekenntnis zur Neugestaltung traditioneller Arbeitsabläufe widerspiegelt (siehe Wirkungsmessung). Andere wiederum haben sich schrittweise Ziele gesetzt, etwa die Ausweitung erfolgreicher Pilotprojekte oder den Aufbau stabiler Plattformen für kontinuierliche Experimente.

Kasten 10. Costa Ricas Contraloría General de la República (CGR): Erleichterung der Einführung

Bottom-up-Innovation

Beim CGR in Costa Rica entwickelte sich eine organische Innovationskultur, in der Mitarbeiter*innen in unterschiedlichen Funktionen dazu ermutigt wurden, KI-Anwendungen zu erforschen. Allen Mitarbeiter*innen stehen KI-Funktionen zur Verfügung, die in die Produktivitätsplattform der Institution integriert sind, was eine umfassende Erprobung ermöglicht hat. Kostenlose Tools werden ebenfalls geprüft, jedoch erst nach Überprüfung durch die IT-Abteilung und den Informationssicherheitsbeauftragten und stets im Einklang mit den institutionellen Richtlinien zur Nutzung künstlicher Intelligenz. Diese Schutzmaßnahmen tragen dazu bei, Risiken im Zusammenhang mit Vertraulichkeit, Datenschutz und verantwortungsvoller Nutzung zu mindern.

Entscheidend ist, dass diese Bottom-up-Kultur in einen umfassenderen Governance-Rahmen eingebettet ist. Das Prüfungsamt hat eine eigene Einheit für Innovation und Lernen in der Aufsicht eingerichtet, die Experimentierinitiativen leitet und das organisationale Lernen steuert. Es wurden formelle Leitlinien für das Management von Experimentierinitiativen herausgegeben, die Grundsätze wie Ko-Kreation, Prototyping und Fehlertoleranz festlegen und gleichzeitig strukturierte Verfahren für Risikobewertung, Validierung und Systematisierung der Ergebnisse vorschreiben. Parallel dazu hat die Institution offizielle Leitlinien für den ethischen und verantwortungsvollen Einsatz von KI eingeführt, die Prinzipien wie Verantwortlichkeit, Transparenz, Fairness, Respekt und Datenschutz festlegen und

zudem konkrete Pflichten für das Personal, die IT-Abteilung sowie den Informationssicherheitsbeauftragten bei der Genehmigung und Überwachung von Tools definieren.

Zusammen gewährleisten diese Rahmenbedingungen, dass Innovationen im CGR in Costa Rica sowohl offen als auch verantwortungsvoll erfolgen: Sie fördern durch Mitarbeiter*innen initiierte Experimente und stellen zugleich ethische Standards, institutionelle Integrität und die Nachhaltigkeit der Einführung von KI sicher.

Unterstützung durch die Leitungsebene

In Costa Rica hat das obere Management eine zentrale Rolle dabei gespielt, Innovation in die Strategie und Kultur der Institution zu integrieren. Der Institutionelle Strategieplan 2021–2026 benennt ausdrücklich *Lernen und Innovation* sowie *Technologie als transformatives Tool* als strategische Treiber, mit klaren Zielen und messbaren Vorgaben in Bezug auf künstliche Intelligenz, datengestützte Aufsicht und organisatorischen Zusammenhalt.

Dieses Bekenntnis auf höchster Ebene wird durch formale Steuerungsinstrumente untermauert. Das Prüfungsamt hat Leitlinien für das Management von Experimentierinitiativen herausgegeben, die einen strukturierten Rahmen für Prototyping, Ko-Kreation und risikobewusstes Innovieren bieten, sowie Leitlinien für den ethischen und verantwortungsvollen Einsatz von KI, die Prinzipien der Transparenz, Sicherheit und institutionellen Verantwortung festlegen. Zusammen gewährleisten diese Instrumente, dass Experimente mit KI sowohl gefördert als auch abgesichert werden.

Die Führung der CGR hat damit die digitale Erkundung legitimiert, Ressourcen bereitgestellt und ein sicheres Umfeld für Mitarbeiter*innen geschaffen, um Innovationen im Einklang mit den strategischen Zielen der Institution zu fördern. Diese Übereinstimmung zwischen Führungsvision, strategischer Planung und operativen Leitlinien hat eine Kultur gefördert, die Kreativität, kalkultiertes Eingehen von Risiken und eine nachhaltige Einführung von KI in der öffentlichen Rechnungsprüfung begünstigt.

Quelle: Contraloría General de la República (CGR) von Costa Rica, 2025

Institutionelle Governance und Aufsicht

Mit dem zunehmenden Umfang und der steigenden Komplexität von KI-Initiativen formalisieren Institutionen verstärkt interne Governance-Mechanismen, um die Koordination der Maßnahmen sicherzustellen und eine strategische Ausrichtung zu gewährleisten. In mehreren Organisationen wurden Räte oder Lenkungsgruppen eingerichtet, um Projektvorschläge zu prüfen, die Umsetzung zu überwachen, Leitlinien und Richtlinien zu entwickeln sowie die Einhaltung ethischer und rechtlicher Vorgaben zu überwachen. Diese Gremien, die häufig aus Vertretern von Leitungsebene, IT, Rechnungsprüfung und Rechtsabteilung bestehen, fungieren sowohl als Befürworter als auch als Garant für Rechenschaftspflicht. Ein Beispiel ist die Einrichtung sowohl eines KI-Rates als auch eines Exekutivrats bei einem SAI, die gemeinsam die Umsetzung der Roadmap und die Ressourcenallokation überwachen. Diese Governance-Strukturen fördern nicht nur Transparenz und Koordination, sondern tragen auch zu einem systematischeren und nachhaltigeren Ansatz für die digitale Transformation bei.

Beim EuRH wurde ein Beratungsausschuss für KI eingerichtet – mit Vertretungen aus den Bereichen Recht, Management, Ethik, Informationssicherheit und Datenschutzbeauftragter –, um die institutionelle Einhaltung des regulatorischen Rahmens für den Einsatz von KI zu erleichtern, die Bewertung und das Management von KI-Risiken zu koordinieren und als interne Fachinstanz für KI-bezogene Fragestellungen zu fungieren.

Neben den Governance-Gremien sind klare und leicht zugängliche interne Richtlinien unerlässlich, um die Lücke zwischen übergeordneten Vorschriften, Berufsstandards und dem täglichen operativen Handeln zu schließen. Sorgfältig ausgearbeitete und auf Befähigung ausgerichtete Leitlinien tragen dazu bei, strategische Ziele und Compliance-Anforderungen in umsetzbare Maßnahmen zu übersetzen und so eine effektive und einheitliche Einführung von KI in der gesamten Institution zu unterstützen. Beispielsweise hat der deutsche BRH eine Leitlinie für die Einführung von Technologien entwickelt, die KI einbeziehen. Dies schließt auch die Datennutzung ein. Die Leitlinie legt den Schwerpunkt auf Prägnanz und Lesbarkeit. Zur Unterstützung der praktischen Umsetzung dieser Leitlinie bietet der BRH zusätzlich KI-Workshops an. In diesen Workshops erwerben die Mitarbeiter*innen praktische und theoretische Grundlagen der KI und machen sich zudem mit dem Compliance-Rahmenwerk des BRH. Bislang haben mehr als 300 Mitarbeiter*innen des BRH an diesen Workshops teilgenommen.

Wirksame Schulungen und Wissensaustausch

Wirksame Schulungsprogramme sind ein entscheidender Bestandteil für die erfolgreiche Einführung und Verbreitung von KI innerhalb der SAls. Die in den Prüfungsteams zu entwickelnden Kompetenzen reichen von technischem Fachwissen und regulatorischem Verständnis bis hin zu Neugier, ethischem Bewusstsein und beruflicher Skepsis. Eine häufige Herausforderung besteht darin, die Lücke zwischen allgemeiner Schulung und praktischer Anwendung zu überbrücken. Wie bei anderem digitalen Wissen kann neu erworbenes Wissen ohne unmittelbare Anwendung schnell wieder in Vergessenheit geraten, wodurch der Wert der Schulung selbst gemindert wird.

Um dem entgegenzuwirken, hat das AGO Singapur ein Modell eingeführt, das die Teilnehmenden verpflichtet, das Gelernte nach Abschluss der Schulung praktisch anzuwenden. Von den Mitarbeiter*innen wird erwartet, ein Projekt oder eine Arbeitsaufgabe mit Bezug zu KI zu initiieren, um das Gelernte zu festigen, die Einführung zu fördern und sicherzustellen, dass die erworbenen Kompetenzen durch praktische Anwendung erhalten bleiben.

Über die technische Schulung hinaus integrieren Institutionen auch Qualifizierungsinitiativen in umfassendere Change-Management-Strategien, die darauf abzielen, eine digitale Kultur zu fördern. In dem Bewusstsein, dass kulturelle Widerstände die Einführung von KI behindern können, investieren einige SAls gezielt in die Einbindung der Führungsebene, strukturierte Kommunikation und weitere Initiativen, um Denkweisen in der gesamten Organisation zu verändern. Mehrere Institutionen haben interne Ausbildungsakademien gegründet oder Partnerschaften mit akademischen Einrichtungen geschlossen, um gezielte Schulungen in den Bereichen KI und Datenwissenschaft anzubieten. Diese Maßnahmen dienen nicht nur dem Wissenstransfer, sondern sollen auch Neugier, digitales Selbstvertrauen und langfristige organisatorische Anpassungsfähigkeit fördern.

Beim EuRH wurde ein Ansatz mit mehreren Kontaktpunkten für Schulungen und Wissensaustausch eingeführt, um die Mitarbeiter*innen mit KI vertraut zu machen. Durch das Angebot verschiedener Beteiligungsformate – von formeller Schulung bis hin zu informellem Austausch im Kollegium – verfolgt diese Strategie das Ziel, organisationsweit Vertrautheit, Vertrauen und praxisbezogenes Verständnis zu fördern (siehe Kasten 11).

Kasten 11. Ein ganzheitlicher Ansatz für KI-Schulungen: Die Erfahrungen des Europäischen Rechnungshofs (EuRH)

Der EuRH verfolgt eine umfassende Strategie zum Aufbau von KI-Kompetenzen bei den Mitarbeiter*innen.

Angesichts des breiten Angebots an allgemeinen KI-Schulungen hat der EuRH seinen Schwerpunkt auf Coaching und gezielte Unterstützung verlagert, um sicherzustellen, dass die Mitarbeiter*innen sowohl theoretisches Verständnis als auch praktische Fähigkeiten für den Einsatz von KI entwickeln.

Um den umfassenden kulturellen Wandel zu unterstützen, der für die Einführung von KI erforderlich ist, hat der Rechnungshof formelle Schulungen mit informellem, kollegialem Lernen und Wissensaustausch kombiniert.

- **Offizielle Schulungen:** Allgemeine KI-Kurse werden sowohl im Rahmen formaler EuRH-Programme als auch über externe Online-Quellen angeboten.
- **GPT-Cafés:** Informelle Zusammenkünfte mit jeweils 20 bis 30 internen Teilnehmenden, die sich auf die Diskussion von KI-Themen konzentrieren, die Teilnehmende eingebracht haben.
- **Knowledge Node:** Ein online verfügbares, kuratiertes Wissensarchiv mit Artikeln, Leitlinien und relevanten KI-Informationen für kontinuierliches, eigenständiges Lernen. Darüber hinaus wird eine Prompt-Bibliothek integriert, die nützliche KI-Prompts zusammenstellt, um die praktische Anwendung von KI zu fördern.
- **Funktionale Foren:** Spezielle Kommunikationskanäle wie ein Funktionspostfach und dedizierte Team-Kanäle, über die Mitarbeiter*innen KI-bezogene Fragen einreichen und zeitnah fachkundige Antworten erhalten können.
- **Konferenzen:** Die Teilnahme an internen und externen Konferenzen ermöglicht einen Einblick in umfassendere Entwicklungen im Bereich der künstlichen Intelligenz sowie in bewährte Verfahren verschiedener Institutionen.

Quelle: Europäischer Rechnungshof, 2025.

Strategische Partnerschaften

Strategische Partnerschaften können auch die institutionellen Fähigkeiten stärken, ohne dass eine Vielzahl technischer Kompetenzen erworben werden muss und ohne die Komplexität, eigene Technologien zu entwickeln und zu pflegen. Einige Organisationen stützen sich auf eine Kombination aus internem Fachwissen und vielfältigen Kooperationen mit der Industrie, der Wissenschaft sowie anderen nationalen oder internationalen Institutionen, um spezialisiertes Wissen zu nutzen und Innovationen zu fördern. Im Bewusstsein der schnellen Obsoleszenz von LLM-Versionen beispielsweise bevorzugen sie die Zusammenarbeit mit externen Anbietern, um maßgeschneiderte Lösungen zu entwickeln, wobei sie gleichzeitig die Eigentumsrechte an den Daten und dem Code behalten und so Innovation und Kontrolle gewährleisten.

Weltweit gilt es als bewährte Praxis, die interne Entwicklung mit strategischen externen Partnerschaften zu kombinieren, insbesondere in den frühen Phasen des Kompetenzaufbaus. Langfristiger Erfolg und Nachhaltigkeit hängen jedoch davon ab, in spezialisierte Teams zu investieren, Entwicklungsprozesse zu formalisieren und KI-Initiativen mit den organisatorischen Prioritäten in Einklang zu bringen. Ohne diese strukturellen Unterstützungsmaßnahmen laufen Institutionen Gefahr, zu stark von externen Akteuren

abhängig zu werden oder nicht in der Lage zu sein, Werkzeuge an sich verändernde Anforderungen anzupassen.

8 Zukünftige Entwicklungen

Mit Blick auf die Zukunft beginnen die fortschrittlichsten Institutionen, die nächste Stufe der KI-Fähigkeiten zu erkunden: „Agentische“ KI-Systeme, die autonom – mit minimalem menschlichem Eingriff – auf definierte Ziele hinarbeiten, kontinuierliche Intelligenz für Echtzeit-Einblicke, Reasoning-Modelle zur Verbesserung der Transparenz sowie multimodale Lösungen, die in der Lage sind, unterschiedliche Datentypen zu verarbeiten.

Obwohl agentische KI in der Praxis noch nicht eingesetzt wird, erforschen Prüfungsinstitutionen aktiv, wie sie dazu beitragen könnte, neue Erkenntnisse zu gewinnen und Risikobereiche aufzudecken, die mit herkömmlichen Methoden nur schwer zu identifizieren sind. Mehrere von ihnen sehen Potenzial in den nächsten drei bis fünf Jahren, während die langfristigen Perspektiven bis zu 20 Jahre reichen. Eine Institution untersucht den Einsatz von KI-Agenten, um Betrugsszenarien zu simulieren und Dokumentenmanipulationen zu erkennen, beispielsweise gefälschte Zertifikate zu identifizieren. Das BAI Korea verfolgt das Ziel, dass KI-Systeme künftig vollständige Prüfberichte verfassen. Diese Experimente werden jedoch mit Vorsicht durchgeführt, da Bedenken hinsichtlich Qualitätssicherung und Vertraulichkeit bestehen. Der norwegische Riksrevisjonen sieht zukünftige Tools vor, die eine Interaktion mit Finanzdaten in natürlicher Sprache oder per Sprachsteuerung ermöglichen sowie die Automatisierung von Managementprüfungen und Qualitätssicherungsprozessen zur Effizienzsteigerung.

Mit dem technologischen Fortschritt sieht die niederländische Algemene Rekenkamer den Einsatz von KI vor, um einen kontinuierlichen Daten- und Informationsfluss zu ermöglichen, sodass Prüfer*innen nahezu in Echtzeit fortlaufend informiert bleiben können. Anstatt sich ausschließlich auf periodische Berichte oder manuelle Datenerhebung zu verlassen, könnten Prüfer*innen tägliche Aktualisierungen erhalten, die für ihren jeweiligen Sektor relevant sind. So würde beispielsweise Prüfer*innen im Gesundheitswesen automatisch über neue Entwicklungen wie Parlamentsdebatten, Medienberichte, regulatorische Änderungen oder aufkommende Risiken informiert werden. Dieses proaktive Modell würde es den Prüfer*innen ermöglichen, stets einen aktuellen und umfassenden Überblick über ihr Fachgebiet zu behalten, wodurch ihre Fähigkeit, Probleme frühzeitig zu erkennen und effektiver darauf zu reagieren, erheblich verbessert würde.

Kurzfristig konzentrieren sich einige Organisationen auf Reasoning-Modelle, die Transparenz schaffen, indem sie aufzeigen, wie Schlussfolgerungen gezogen wurden, und so möglicherweise die Anforderungen an ein Prüfprotokoll unterstützen. Diese Modelle könnten die Begründung hinter den Ergebnissen dokumentieren und KI-generierte Erkenntnisse nachvollziehbarer machen, wodurch sie potenziell näher an KI-Ausgaben heranrücken, die den Prüfungsstandards entsprechen. Neue Anwendungsfälle betreffen auch die Integration multimodaler Modelle (die Informationen aus verschiedenen Datentypen verarbeiten und integrieren können), darunter Werkzeuge, die Informationen aus qualitativ minderwertigen Bildern extrahieren können, wie beispielsweise gescannte oder beschädigte Abbildungen von Geburtsurkunden. Diese Fähigkeiten können den Anwendungsbereich von KI in dokumentenintensiven Prüfungsumgebungen erheblich erweitern.

Weitere zukünftige Anwendungsbereiche umfassen fortschrittlichere prädiktive Analysen, beispielsweise die Entwicklung von Modellen zur Erkennung von Abweichungen der Beschaffungspreise von den Marktpreisen. Andere Institutionen richten ihren Fokus zunächst nach innen und erkennen die

Notwendigkeit, die bereits verfügbaren großen Datenmengen besser zu nutzen, bevor sie sich auf komplexere KI-Lösungen ausweiten.

Mit Blick auf die Zukunft könnten die Weiterentwicklung analytischer Fähigkeiten und das Aufkommen fortschrittlicherer KI-Ansätze, wie agentische KI, das Potenzial haben, ganze Prozesse und damit verbundene Entscheidungsfindungen zu automatisieren. Auch wenn die Einführung mit Vorsicht erfolgen wird, könnte dieser Wandel eine grundlegende Überprüfung der derzeitigen Prüfmethoden und -standards sowie Anpassungen der einschlägigen Vorschriften und Rechtsrahmen erfordern, um die Gültigkeit und Verlässlichkeit der Prüfungsfeststellungen weiterhin zu gewährleisten. Die rasante Entwicklung dieser Technologien unterstreicht zusätzlich die Notwendigkeit von Schlüsselkompetenzen und einer anpassungsfähigen digitalen Kultur in allen Institutionen.

Parallel zu diesen Entwicklungen verändert sich auch der breitere Kontext des öffentlichen Sektors: Zahlreiche zentrale Regierungsstellen, darunter Finanzministerien und Finanzagenturen, bauen zunehmend eigene KI-Kompetenzen auf. Dies eröffnet potenzielle Möglichkeiten für eine engere Zusammenarbeit zwischen Prüfungsinstitutionen und anderen öffentlichen Verwaltungen im Bereich der KI. Eine solche Zusammenarbeit könnte dazu beitragen, Doppelarbeit zu vermeiden, prüfungsrelevante Anforderungen wie Nachvollziehbarkeit, Erklärbarkeit und Governance bereits in der frühen Phase der Entwicklung von KI-Systemen zu berücksichtigen und so zur Bündelung gemeinsamen Fachwissens beizutragen. Auch wenn dies noch keine weit verbreitete Praxis ist, deutet es auf eine mögliche zukünftige Entwicklung innerhalb eines zunehmend vernetzten KI-Ökosystems im öffentlichen Sektor hin.

9 Schlussfolgerungen

Die Integration von KI-gestützten Methoden kann im Bereich der öffentlichen Rechnungsprüfung erhebliche Verbesserungen hinsichtlich Produktivität (Effizienz und Wirksamkeit), Reaktionsfähigkeit und Rechenschaftspflicht bewirken. Fortgeschrittene Modelle ermöglichen die automatische Erkennung von Anomalien in Finanzdaten und unterstützen vorbeugende Maßnahmen, indem sie auf risikobehaftete Situationen hinweisen. KI kann zudem das Wissensmanagement von Prüfer*innen verbessern, indem sie die Extraktion von Schlüsselinformationen aus großen Mengen an Textdokumenten erleichtert und die Erstellung strukturierter Datensätze aus unstrukturierten Quellen ermöglicht. Darüber hinaus können LLMs die Erstellung, Zusammenfassung und Übersetzung von technischem Material unterstützen und so eine höhere Konsistenz gewährleisten, während Computer Vision und Deep-Learning-Modelle die Bewertung physischer Infrastrukturen durch die Analyse von visuellen Daten ermöglichen.

Die institutionsübergreifende Analyse zeigte, dass das transformative Potenzial von KI in der Rechnungsprüfung ebenso stark von der zugrunde liegenden Hard- und Software-Infrastruktur abhängt wie von den Prozessen des Datenmanagements und den Entwicklungskapazitäten. Die Auswahl von Hard- und Software offenbart eine Abwägung zwischen der anfänglichen Investition in autonome Lösungen und der Abhängigkeit von externen Anbietern. Dies spiegelt Überlegungen zu Datensicherheit, Compliance und Anbieterabhängigkeit wider. Während fragmentierte Datenquellen und heterogene Architekturen weiterhin erhebliche Engpässe darstellen, können durch Investitionen in Vorverarbeitungstools und klarere Governance-Rahmenbedingungen eine höhere Datenqualität und Standardisierung erreicht werden. Die Entwicklungskapazitäten variieren erheblich, und Institutionen, die interne Teams für die Entwicklung maßgeschneiderter KI-Lösungen aufgebaut haben, scheinen langfristig besser aufgestellt zu sein, um einen erfolgreichen und nachhaltigen KI-Übergang zu erreichen.

Strategische Überlegungen gehen über die Infrastruktur hinaus. Das Fehlen von Prüfungsstandards für den Einsatz von KI in der Prüfung führt zu Unsicherheiten, während ethische und rechtliche Bedenken hinsichtlich Erklärbarkeit, Transparenz und Verzerrungen die Notwendigkeit menschlicher Aufsicht unterstreichen. In diesem Zusammenhang spielen Organisationskultur und praxisnahe Schulungen eine entscheidende Rolle dabei, die Nutzer*innen dazu zu befähigen, KI-Ergebnisse kritisch zu bewerten und ihre aktive Mitwirkung am Entwicklungsprozess zu fördern. Erfolgreiche Beispiele zeigen auch, dass Innovationen aus der Belegschaft, wenn sie durch eine starke Governance und das Engagement des Managements unterstützt werden, die Einführung erheblich fördern und sicherstellen können, dass KI-Lösungen sowohl wirksam als auch rechenschaftspflichtig bleiben.

Bisher stellt KI eine wirkungsvolle Ergänzung zu traditionellen Prüfmethoden dar, ersetzt jedoch nicht das professionelle Urteilsvermögen. Für eine wirkungsvolle Nutzung sind ganzheitliche Strategien erforderlich, die Infrastrukturkapazitäten, Governance-Rahmen und die Qualifizierung des Personals miteinander verbinden. Durch die Verbesserung der Datenqualität, den Aufbau anpassungsfähiger Kapazitäten und die Förderung einer innovationsfreundlichen Kultur können Prüfungsinstitutionen nicht nur ihre operative Effektivität steigern, sondern auch das öffentliche Vertrauen in die Integrität und Rechenschaftspflicht der Regierung stärken.

Der zunehmende Einsatz von KI-Tools hat erhebliche Auswirkungen auf die internen Kontrollsysteme innerhalb von Prüfungsinstitutionen. Da KI zunehmend in die Risikobewertung, Datenanalyse und Beweiserhebung integriert wird, werden diese Werkzeuge faktisch zu einem Bestandteil des

Prüfungsprozesses und müssen daher denselben Kontrollprinzipien unterliegen wie andere kritische Prüfungssysteme. Eine starke interne Kontrollumgebung, einschließlich klarer Prozessverantwortlichkeiten, Zuständigkeiten, robuster Daten-Governance, Zugangskontrollen, Validierungsverfahren, Dokumentation, Prüfpfade usw., ist eine Voraussetzung für den zuverlässigen und verantwortungsvollen Einsatz von KI. Wo interne Kontrollen schwach oder fragmentiert sind, wird KI Fehler und Unstimmigkeiten eher verstärken als die Prüfungsqualität verbessern. Die Einbettung von KI in ein solides System interner Kontrolle ist daher unerlässlich, um sicherzustellen, dass technologische Innovation die Glaubwürdigkeit, Rechenschaftspflicht und das öffentliche Vertrauen in die Prüfung stärkt und nicht untergräbt.

Über ihre interne Nutzung von KI hinaus sehen sich Prüfungsinstitutionen zunehmend mit der übergeordneten Frage konfrontiert, wie KI im öffentlichen Sektor entwickelt, implementiert und gesteuert wird. Während Regierungen KI in zentrale Politikbereiche, das Finanzmanagement und die Dienstleistungserbringung integrieren, werden SAls voraussichtlich eine zentrale Rolle dabei spielen, zu prüfen, wie Verantwortlichkeit, Transparenz und Kontrollmechanismen in durch KI unterstützten Entscheidungsprozessen sichergestellt werden. Dies versetzt die SAls nicht nur in die Lage, KI-gestützte Systeme zu prüfen, sondern auch dazu beizutragen, die öffentliche Debatte über Verantwortlichkeit, Nachvollziehbarkeit und Aufsicht zu gestalten – gestützt auf ihr Mandat und ihre Erfahrung bei der Bewertung von Governance-Strukturen. In diesem sich wandelnden Kontext könnte die Prüfung von KI ebenso sehr die Überprüfung institutioneller Rahmenbedingungen und Schutzmechanismen betreffen wie die Bewertung einzelner Technologien.

Glossar

Die folgenden Definitionen dienen dem besseren Verständnis der Lesenden und stellen, sofern nicht anders angegeben, keine offiziellen Definitionen der OECD dar.

Application Programming Interface (API): bezeichnet eine Sammlung von Routinen, Protokollen und Werkzeugen zur Entwicklung von Softwareanwendungen. Eine API legt fest, wie Softwarekomponenten miteinander interagieren sollten, um Daten, Funktionen und Funktionalitäten auszutauschen. Eine gute API erleichtert die Entwicklung eines Programms, indem sie alle Bausteine bereitstellt. (Quelle: Europäische Kommission)

Cloud-native Dienste: Dienste, die den Nutzer*innen auf Abruf über das Internet von den Servern eines Cloud-Computing-Anbieters bereitgestellt werden, anstatt von den eigenen lokalen Servern eines Unternehmens. (Quelle: Europäische Kommission)

Convolutional Neural Networks (CNNs): eine Art von Deep-Learning-Algorithmus, der für die Verarbeitung und Analyse visueller Daten wie Bildern und Videos entwickelt wurde. CNNs verwenden Schichten miteinander verbundener Knoten, die Faltungsoperationen ausführen, um Muster wie Kanten, Texturen und Formen zu erkennen und zu erlernen, wodurch sie sich besonders für Aufgaben wie Bilderkennung, Klassifizierung und Computer Vision eignen. (Quelle: IBM)

Crawler: ein automatisiertes Programm, das das Internet nach neuen Web-Dokumenten durchsucht und deren Adressen sowie inhaltsbezogene Informationen in einer Datenbank ablegt, auf die über eine Suchmaschine zugegriffen werden kann. (Quelle: Europäische Kommission)

Individuelle Ensemble-Methode: Verfahren im maschinellen Lernen, bei denen mehrere Modelle kombiniert werden, um die Vorhersagegenauigkeit zu erhöhen, Fehler zu reduzieren und die Robustheit zu steigern. (Quelle: IBM)

Domänenspezifisches LLM: Bezeichnet ein großes Sprachmodell (LLM), das für ein bestimmtes Fachgebiet oder Themenfeld trainiert oder feinabgestimmt wurde, um bei domänenspezifischen Aufgaben bessere Leistungen zu erzielen als ein allgemeines LLM. (Quelle: IBM)

Extract-transform-load (ETL): ein Datenverarbeitungs-Workflow, bei dem Daten aus unterschiedlichen Quellen extrahiert, in ein geeignetes Format transformiert und in ein Zielsystem, wie beispielsweise ein Data Warehouse, geladen werden. (Quelle: IBM)

Graphenbasiertes Modell: eine Datenstruktur oder ein Algorithmus, der Informationen als Knoten (Entitäten) und Kanten (Beziehungen) darstellt und für die Analyse komplexer Netzwerke wie sozialer Verbindungen oder Wissensgraphen verwendet wird.

Graphics Processing Unit (GPU): bezeichnet eine elektronische Schaltung, die speziell für die digitale Bildverarbeitung entwickelt wurde und die Computergrafik beschleunigt. Die Fähigkeit von GPUs, in kurzer Zeit eine enorme Anzahl an Berechnungen auszuführen, hat sie im Bereich der künstlichen Intelligenz

unentbehrlich gemacht, da sie dort zur Bewältigung datenintensiver und rechenaufwendiger Aufgaben eingesetzt werden. (Quelle: IBM)

Intelligent Document Processing (IDP): bezeichnet den Einsatz von KI, um automatisch strukturierte Daten aus unstrukturierten oder semi-strukturierten Quellen wie PDFs, gescannten Dokumenten, Bildern und handschriftlichen Formularen zu extrahieren. [Definition aus dem Bericht]

Internet of Things (IoT): Internet der Dinge, ein Netzwerk miteinander verbundener physischer Geräte, die mit Sensoren und Software ausgestattet sind und so in der Lage sind, Daten zu erfassen, auszutauschen und darauf zu reagieren. (Quelle: Oracle)

Intersection over Union (IoU): auch als Jaccard-Index bekannt, misst, wie stark sich zwei Formen, die dasselbe Objekt in einem Bild lokalisieren, überlappen, indem die Flächen der vorhergesagten Box und der tatsächlichen Box verglichen werden. Ein höherer IoU-Wert entspricht einer größeren Überlappung und somit einer besseren Übereinstimmung. IoU ist die gebräuchlichste Bewertungskennzahl für Aufgaben wie Segmentierung, Objekterkennung und -verfolgung. (Quelle: Stanford University)

Gekennzeichnete Daten: Datenpunkte, die mit einem oder mehreren Labels versehen wurden, die deren Bedeutung, Kategorie oder Kontext angeben. Diese Labels dienen während des Modelltrainings als Referenzergebnisse und ermöglichen es den Modellen, zu lernen, indem sie ihre Vorhersagen mit den zugewiesenen Labels vergleichen.

Große Sprachmodelle (LLMs): bezeichnen eine Klasse von künstlichen Intelligenzmodellen, die auf einer großen Menge an Textdaten trainiert werden, um menschliche Sprache zu verstehen und zu generieren. (Quelle: Stanford University)

Maschinelles Lernen (ML): ist eine Reihe von rechnergestützten Techniken, die es Maschinen ermöglicht, auf automatisierte Weise durch Muster und Schlussfolgerungen zu lernen, anstatt durch explizite Anweisungen eines Menschen. Beim Maschinellen Lernen werden Maschinen häufig darauf trainiert, ein Ergebnis zu erzielen, indem ihnen zahlreiche Beispiele für korrekte Ergebnisse gezeigt werden. Sie können jedoch auch eine Reihe von Regeln festlegen und die Maschine durch Versuch und Irrtum lernen lassen. (Quelle: OECD)

Modelldrift: bezeichnet die Verschlechterung der Leistungsfähigkeit eines Machine-Learning-Modells aufgrund von Veränderungen in den Daten oder in den Beziehungen zwischen Eingabe- und Ausgabewerten. Modelldrift – auch als Modellverfall bekannt – kann die Modellleistung negativ beeinflussen und zu fehlerhaften Entscheidungen sowie schlechten Prognosen führen. (Quelle: IBM)

Natural Language Processing (NLP): das Teilgebiet der künstlichen Intelligenz, das darauf abzielt, Computern das Verstehen, Interpretieren und Erzeugen menschlicher Sprache auf sinnvolle Weise zu ermöglichen. (Quelle: IBM)

Overfitting: Im Maschinellen Lernen tritt Overfitting auf, wenn ein Algorithmus sich zu stark oder sogar exakt an die Trainingsdaten anpasst, sodass das Modell keine genauen Vorhersagen oder Schlussfolgerungen für andere als die Trainingsdaten treffen kann. (Quelle: IBM)

Optische Zeichenerkennung (OCR): eine Technologie, die verschiedene Arten von gedrucktem oder handschriftlichem Text aus Bildern oder gescannten Dokumenten in maschinenlesbaren und bearbeitbaren digitalen Text umwandelt. (Quelle: IBM)

Production-grade: Bezieht sich auf Software oder Systeme, die zuverlässig, skalierbar und robust genug sind, um in realen Betriebsumgebungen eingesetzt und genutzt zu werden. (Quelle: IBM)

Retrieval-Augmented Generation (RAG): eine KI-Technik, die es einem generativen Modell ermöglicht, Informationen aus gezielten Datenquellen abzurufen, um Antworten zu erzeugen, die auf einem festgelegten Dokumentensatz basieren. (Quelle: IBM)

Robotic Process Automation (RPA): ist eine Technologie, die Software-„Roboter“ einsetzt, um repetitive, regelbasierte digitale Aufgaben zu automatisieren. Sie imitiert menschliche Interaktionen mit Systemen und Anwendungen, um Prozesse schneller und mit weniger Fehlern abzuschließen, wodurch die Effizienz gesteigert und Mitarbeiter*innen in die Lage versetzt werden, sich auf höherwertige Aufgaben zu konzentrieren. (Quelle: IBM)

Abgeschlossene Umgebungen: isolierte Computerumgebungen, die sämtliche erforderliche Software und Abhängigkeiten enthalten und so eine konsistente und reproduzierbare Leistung über verschiedene Systeme hinweg gewährleisten. (Quelle: IBM)

Semantische Recherche: ist eine Datensuchtechnik, die darauf abzielt, die kontextuelle Bedeutung und Absicht hinter einer Suchanfrage von Nutzer*innen zu verstehen, anstatt lediglich Schlüsselwörter abzugleichen. (Quelle: Google)

Spark-Cluster: eine Gruppe miteinander verbundener Computer, die gemeinsam große Datenmengen mithilfe von Apache Spark, einem schnellen und universell einsetzbaren Cluster-Computing-System, verarbeiten. (Quelle: Apache Spark)

Überwachtes maschinelles Lernen: eine Form des maschinellen Lernens, bei der Modelle mit gekennzeichneten Daten trainiert werden und auf Basis von Eingabe-Ausgabe-Beispielen lernen, Ergebnisse vorherzusagen. (Quelle: IBM)

Unüberwachtes maschinelles Lernen: Ein Ansatz des maschinellen Lernens, der Muster oder Gruppierungen in Daten ohne vorgegebene Ergebnisse identifiziert und häufig zur Erkennung von Anomalien eingesetzt wird. (Quelle: IBM)

Vektor-Datenbanken: eine Datenbank, die für die Speicherung und Suche von Daten in Form von Vektoren (numerische Arrays) optimiert ist und häufig für Ähnlichkeitssuchen in KI-Anwendungen wie Empfehlungssystemen verwendet wird. Vektordatenbanken spielen eine entscheidende Rolle bei der Ermöglichung fortschrittlicher KI-Anwendungen, indem sie hochdimensionale Datenrepräsentationen – sogenannte Vektoren oder Embeddings – effizient speichern und durchsuchen, wobei KI-Modelle wie große Sprachmodelle Texte, Bilder oder andere Daten in diese Vektoren umwandeln, um deren Bedeutung zu erfassen. (Quelle: Pinecone)

Literaturverzeichnis

- (n.a.) (2025), *The Agentic AI landscape: Conceptual foundations, key features and illustrative examples*, [https://one.oecd.org/official-document/DSTI/DPC/GPAI\(2025\)18/en](https://one.oecd.org/official-document/DSTI/DPC/GPAI(2025)18/en) (accessed on 13 January 2026). [8]
- CGU (2025), <https://www.gov.br/ancine/pt-br/assuntos/atribuicoes-ancine/fomento/prestacao-de-contas/malha-fina>. [3]
- Huang, Q. et al. (2024), “A graph says more than a thousand journal entries: Harnessing graph autoencoder networks in auditing”, https://www.efk.admin.ch/wp-content/uploads/publikationen/fachtexte/2024_12_a_graph_says_more_than_a_thousand_journal_entries.pdf (accessed on 30 October 2025). [2]
- OECD (2025), *Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions*, OECD Publishing, Paris, <https://doi.org/10.1787/795de142-en>. [1]
- OECD (2024), “Governing with Artificial Intelligence: Are governments ready?”, *OECD Artificial Intelligence Papers*, No. 20, OECD Publishing, Paris, <https://doi.org/10.1787/26324bc2-en>. [7]
- OECD (2024), *Recommendation of the Council on Artificial Intelligence*, <https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449>. [6]
- OECD (2019), *The Path to Becoming a Data-Driven Public Sector*, OECD Digital Government Studies, OECD Publishing, Paris, <https://doi.org/10.1787/059814a7-en>. [4]
- OECD (forthcoming), *Governing with Artificial Intelligence: Measuring impact and demonstrating returns on AI investments*. [5]