RAG-Chatbots im Kundenservice von Unternehmen und der öffentlichen Verwaltung

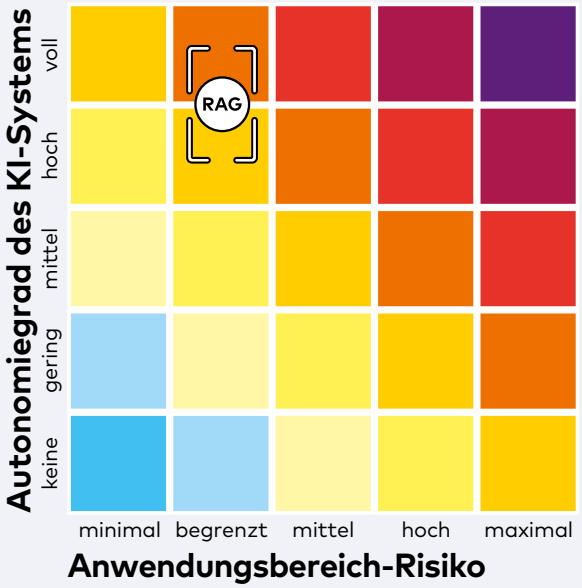
Executive Summary

Unser Leben verlagert sich zunehmend in den digitalen Raum. Die vermehrte digitale Kommunikation mit Kund:innen und Bürger:innen stellt Unternehmen und die öffentliche Verwaltung vor Herausforderungen, denn eine rasche und effektive Bearbeitung von Anfragen und Anliegen ist auch mit hohen Kosten verbunden. Chatbots mit künstlicher Intelligenz (KI), insbesondere sogenannte Retrieval-Augmented-Generation-(RAG)-Chatbots, gewinnen daher zunehmend an Bedeutung, da sie eine große Anzahl von Anfragen schnell und kosteneffizient bearbeiten können.

Die bisherigen praktischen Erfahrungen verschiedener Länder beim Einsatz von RAG-Systemen zeigen, dass sich damit sowohl für Unternehmen als auch für Nutzer:innen erhebliche Vorteile erzielen lassen können.[2,6] RAG-Chatbots rufen Informationen aus internen oder externen Wissensquellen ab, die nicht im Trainingsdatensatz der Sprachmodelle (LLM) enthalten sind. Nutzer:innen erhalten dadurch genauere Antworten. Organisationen können durch ihren Einsatz Kosteneffizienz und das Vertrauen der Nutzer:innen gewinnen. Allerdings gibt es auch erhebliche Herausforderungen, die diesen Vorteilen gegenüberstehen. Aufgrund der eingesetzten Grundlagenmodelle (LLMs) sind auch RAG-Chatbots bei mangelnder Datenqualität und bei hoher Komplexität von Abfragen anfällig für sog. Halluzinationen, also Falschaussagen. Solche Fehlinformationen sind besonders problematisch, wenn sie von Behörden stammen, da Bürger:innen deren Auskünfte als verlässlich ansehen. Insgesamt können Fehlinformationen schwerwiegende Folgen, etwa rechtswidriges Verhalten, Verlust von Rechtsansprüchen oder die Erosion des Vertrauens in öffentliche Institutionen, nach sich ziehen.

Insgesamt geht mit der Nutzung von RAG-Chatbots lt. ALAIT Risikoradar ein „hohes Risiko" (dunkelorange, siehe Grafik) einher. Zwar wird das Risiko bei der Nutzung von RAG-Systemen im Kundenservice von Unternehmen und der öffentlichen Hand gemäß *EU AI-Act* als „begrenzt" (Stufe 2) eingestuft, weil anzunehmen ist, dass sie – anders als bei automatisierten Entscheidungen z.B. über staatliche Transferleistungen – keine unmittelbar weitreichenden Folgen für auskunftssuchende Personen haben. In Bezug auf den *Autonomiegrad* sind die zur Anwendung kommenden RAG-Systeme jedoch in der Regel meist voll autonom, da sie unabhängig arbeiten und nur minimale oder gar keine menschliche Intervention erfordern. Die Beteiligung des Menschen beschränkt sich somit auf die langfristige Aufsicht mittels Audits (Stufe 5).

ALAIT Risikoradar für RAG-Chatbots im Kundenservice von Unternehmen und der öffentlichen Verwaltung



RAG Retrieval-Augmented-Generation (RAG) Chatbots

Das ALAIT Risikoradar ist ein wissenschaftlich entwickeltes Risikoanalysetool für Künstliche Intelligenz (KI), das KI-Anwendungen kontextbezogen und unter Berücksichtigung ihres technischen Autonomiegrades einstuft und so die Risikosphäre für Anwender:innen auf einen Blick sichtbar macht. Dabei gilt: Je höher das Einsatz-Risiko aus dem Anwendungskontext und je größer der Autonomiegrad des KI-Systems in Bezug auf Entscheidungen, desto riskanter ist der Einsatz einzustufen. Eine erweiterte Klammer weist auf eine Bandbreite in der Risikoeinstufung hin. Ein geringer Autonomiegrad eines KI-Systems bedeutet nicht, dass man sich zurücklehnen kann. Es erfordert eine starke Rolle der Menschen, die es anwenden. (Details zum Stufenmodell s. S. 8f).

**ALAIT ist ein vom Bundesministerium für Innovation, Mobilität und Infrastruktur (BMIMI) initiiertes Forschungsprojekt zur Entwicklung einer neuartigen Form von Technologiebewertung für KI-Technologien mit wissenschaftlichen Methoden.*

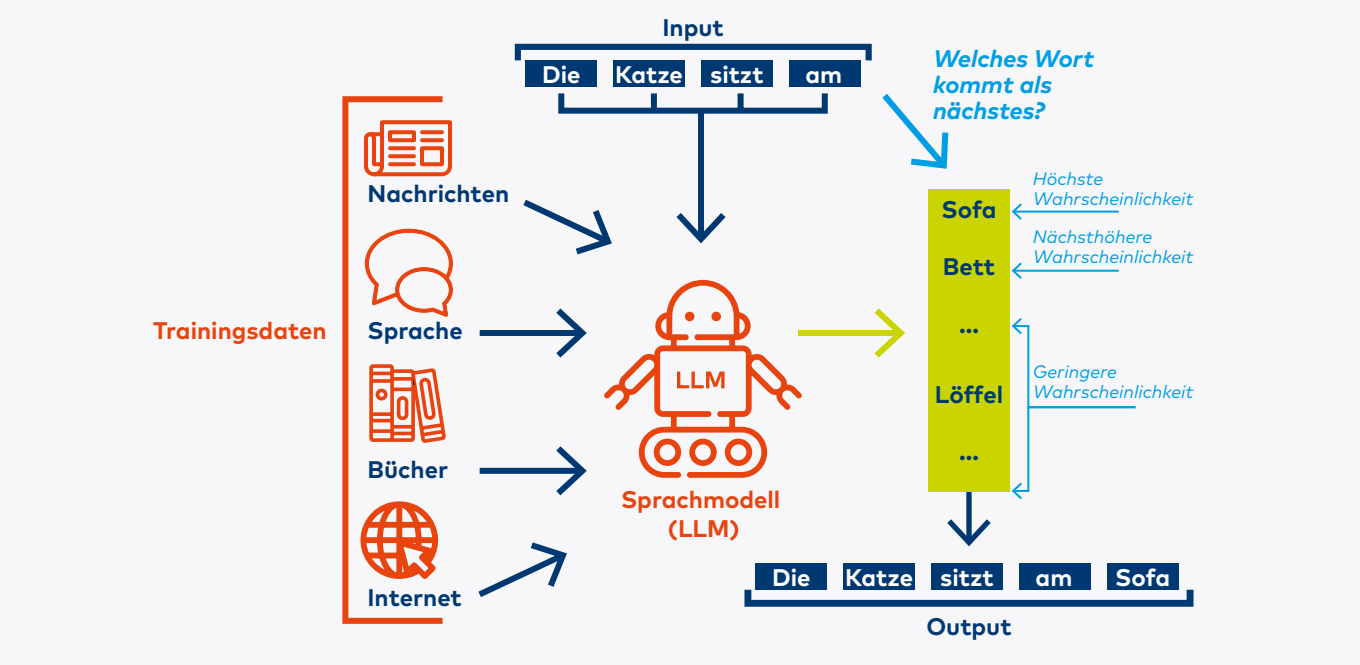
Organisationen und Systementwickler:innen können das Risiko reduzieren, indem sie Maßnahmen zur Risikominderung umsetzen. Grundsätzlich sind die Richtlinien der Datenschutz-Grundverordnung (DSGVO) und des EU AI-Acts, die unter anderem auch die Transparenzrichtlinien enthalten, einzuhalten. Darüber hinaus können Probleme mit dem Datenschutz vermieden werden, indem präzise und hochwertige Daten verwendet werden, die keine sensiblen (personenbezogenen) Daten beinhalten. Die menschliche Aufsicht von RAG-Systemen in Form von Kontrollen der Ergebnisse und Sicherheitslücken ist unverzichtbar. Nutzer:innen von RAG-Chatbots müssen in jedem Fall über die Verwendung ihrer Daten informiert werden.

Einleitung und Technologiebeschreibung

Die Technische Universität München (TUM) hat im Jahr 2024 das von einem ihrer Start-ups entwickelte RAG-System „OneTutor" eingeführt, um ihren Studierenden tagesaktuelle Informationen zu Vorlesungen und Studienfächern der Universität zur Verfügung zu stellen. Das erfolgreiche Informationssystem wird heute bereits von rund 9.700 Studierenden an 22 Universitäten genutzt.[9]

Wie die TUM setzen nun auch vermehrt Unternehmen und die öffentliche Verwaltung auf RAG-Chatbots, um mit der hohen Anzahl an Anfragen von Kund:innen und Bürger:innen Schritt halten zu können.

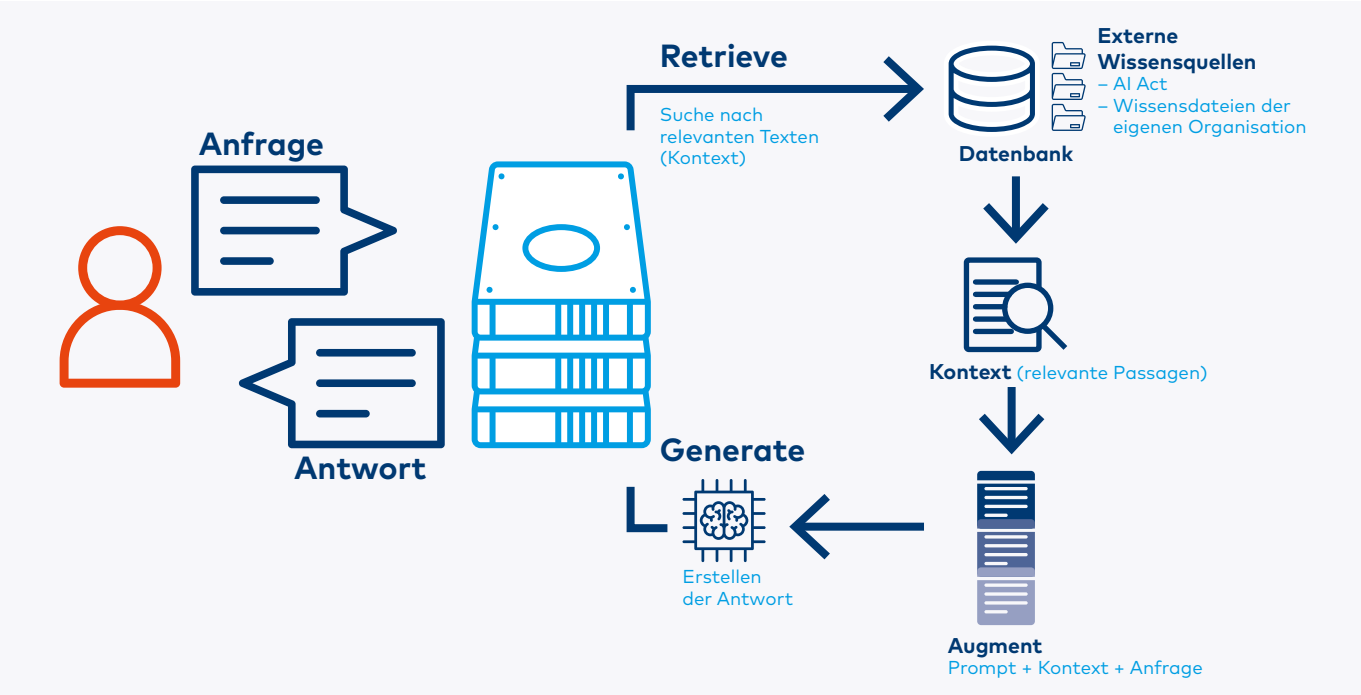
RAG-Chatbots funktionieren auf Basis von Large Language Models (LLMs), auch Sprachmodelle genannt. Sie sind die Haupt-KI-Komponente aller RAG-Systeme. LLMs nehmen die Benutzereingabe entgegen und erstellen auf Grundlage der Informationen, mit denen sie trainiert wurden, eine Antwort. Das folgende Diagramm zeigt das Grundprinzip von Sprachmodellen (LLMs):



Ein Sprachmodell (LLM) lernt, indem es aus sehr vielen Textsorten wie Nachrichten, Büchern, Gesprächen und Internetseiten Muster erkennt, wie Wörter und Sätze typischerweise zusammenhängen. Wenn man dem Modell einen Satzanfang wie „Die Katze sitzt am..." vorgibt, berechnet es das mit der höchsten Wahrscheinlichkeit folgende Wort. Dabei prüft es zahlreiche Möglichkeiten, etwa „Sofa", „Bett" oder „Löffel", und wählt schließlich das wahrscheinlichste Wort aus. So entsteht der vollständige Satz „Die Katze sitzt am Sofa". Kurz gesagt ist ein LLM ein Programm, das auf Basis von gelerntem Sprachwissen vorhersagt, welches Wort als Nächstes am besten passt.

RAG-Chatbots rufen Informationen aus internen oder externen Wissensquellen ab, die nicht im Trainingsdatensatz des LLMs enthalten sind. Wenn eine Person eine Frage stellt, sucht das System zuerst in diesen Datenquellen nach passenden Informationen und nutzt dann ein LLM, um eine Antwort zu erzeugen.[2]

Das folgende Diagramm zeigt das Grundprinzip von RAG-Chatbots:[13,23]



Ein RAG-Modell arbeitet in folgenden Schritten:[3,4,5]

1. **Frage umwandeln:** Die eingegebene Frage wird in eine spezielle Zahlenform (Vektor) übersetzt. Damit führt das System eine semantische Suche durch.

2. **Informationen abrufen:** Danach sucht das System nach relevanten Informationen – entweder in den eigenen Daten oder in vordefinierten Quellen wie spezifischen Datenbanken oder Dokumenten. Die gefundenen Texte werden ebenfalls in Vektoren umgewandelt und in einer sogenannten Vektordatenbank gespeichert.

3. **Antwort erzeugen:** Anschließend erstellt ein Sprachmodell (LLM) auf Basis der bereitgestellten Informationen die Antwort. In fortgeschritteneren Varianten kann das System die Antwort nochmals mit der ursprünglichen Frage vergleichen, um sie zu verbessern oder zu präzisieren.

So entsteht eine Antwort, die auf aktuelleren und relevanteren Informationen basiert als bei herkömmlichen Chatbots (z.B. ChatGPT oder Gemini). Dadurch können RAG-Systeme auch Fragen beantworten, zu denen die Sprachmodelle selbst kein direktes Wissen aus dem Training haben.

Trotz erfolgreicher Anwendungsbeispiele ist die Anwendung von RAG-Chatbots in der Realität mit Herausforderungen verbunden. So etwa lieferte der generative KI-Chatbot der Stadt New York (MyCity Chatbot) eine Reihe von Antworten, die die Nutzer:innen entweder zu Gesetzesverstößen ermutigten oder unrichtige Informationen über das Gesetz enthielten.[11] Der Chatbot gab etwa die Auskunft, dass es legal sei, Personen zu entlassen, die sich über sexuelle Belästigung beschwert hatten.[11,24]

Auch im Unternehmenskontext können die Folgen fehlerhafter Auskünfte von RAG-Chatbots gravierend sein. So gab der von Air Canada eingesetzte Chatbot einem Kunden falsche Informationen zur Tarifpolitik der Fluggesellschaft im Trauerfall. Der Kunde verließ sich auf diese Informationen und klagte daraufhin erfolgreich auf Rückerstattung. Der Fall führte zu rechtlichen Konsequenzen für das Unternehmen und zu einem erheblichen Vertrauensverlust sowie Reputationsschäden.[11,25] Solche Beispiele verdeutlichen, dass unzureichend geprüfte oder nicht regelmäßig aktualisierte RAG-Systeme sowohl im öffentlichen als auch im privaten Sektor erhebliche Risiken in Bezug auf Haftung, Verbraucherschutz und institutionelles Vertrauen darstellen.

Im Folgenden werden die wichtigsten Chancen und Herausforderungen von RAG Chatbots noch einmal systematisch aufgelistet.

Chancen

Der Einsatz von RAG-Chatbots in der öffentlichen Verwaltung bietet gegenüber herkömmlichen Chatbots eine Reihe von Vorteilen für Endnutzer:innen und einsetzende Organisationen:

- **Geschwindigkeit und *Genauigkeit:*** Obwohl die Informationen in mehreren Verarbeitungsstufen produziert werden, sind RAG-Chatbots schneller und deutlich genauer als herkömmliche Chatbots.[1,3,6,31] Wie eine Studie von Deloitte in Kanada gezeigt hat, können durch die Automatisierung wiederkehrender Aufgaben (mit Hilfe von KI-Tools wie RAG-Chatbots) zudem Verarbeitungsfehler von Sachbearbeiter:innen um bis zu 40 % reduziert werden.[7] Diese Verbesserungen sollten jedoch mit Vorsicht betrachtet werden, da sie stark von der Datengrundlage, dem Anwendungsbereich und der Evaluierungsmethode abhängen.

- **Vereinfachter Zugang zu öffentlichen Informationen:** RAG-Chatbots können rund um die Uhr Informationen aus komplexen rechtlichen oder administrativen Quellen wie Gesetzen, Formularen und Zulassungskriterien abrufen und bürokratische Sprache in einfache, verständliche Begriffe übersetzen. Dadurch werden Hindernisse für Bürger:innen abgebaut, denen der Umgang mit oft komplizierten öffentlichen Dokumenten schwerfällt.[11,21] Ein Beispiel dafür ist die Rundfunk und Telekom Regulierungs-GmbH (RTR), die ein RAG-System entwickelt hat, das Fragen zum EU AI-Act und zu regulatorischen Rahmenbedingungen zu KI effektiv und ***transparent*** beantworten kann, weil die genaue Informationsquelle angezeigt werden kann.[23]

- **Vertrauen der Nutzer:innen:** Durch die Echtzeitverbindung der RAG-Chatbots mit den Datenbanken können sie aktuelle und kontextbezogene Informationen liefern und das Interaktionserlebnis der Nutzer:innen[1,8,10] und somit die allgemeine Zufriedenheit mit dem Service im Vergleich zu herkömmlichen Chatbots verbessern.[2,5]

- **Arbeitsentlastung:** RAG-Chatbots können sowohl standardisierte als auch nicht standardisierte, häufig wiederkehrende Anfragen automatisiert beantworten und relevante Informationen aus geprüften Dokumenten abrufen. Dadurch wird der Aufwand für Mitarbeiter:innen bei der Routinekommunikation deutlich reduziert. Die frei werdende Zeit kann für komplexe Fälle, individuelle Beratung oder Entscheidungen, die menschliches Urteilsvermögen erfordern, verwendet werden.[31]

- **Aktuelle Informationen:** RAG-Systeme können das LLM nicht nur mit lokalen Datenbanken, sondern auch direkt mit Nachrichtenseiten oder anderen häufig aktualisierten Informationsquellen verbinden. Dadurch werden sowohl Mitarbeiter:innen als auch Nutzer:innen die relevantesten Informationen bereitgestellt.[2,5]

- **Kundenzufriedenheit:** Durch die Integration von RAG-Chatbots erhalten Kund:innen eine kontextbezogene Unterstützung. Sie erhalten sofortigen Kundenservice mit personalisierter, mehrsprachiger und barrierefreier Kommunikation – etwas, das bei menschlichem Kundenservice nicht immer möglich ist.[10]

- **Kosteneffizienz:** Unternehmen wie Vodafone, Alibaba und Klarna geben an, seit der Integration von KI-gestützten Chatbots Kosteneinsparungen in Millionenhöhe erwirtschaftet zu haben.[8] Das Aufsetzen unternehmensspezifischer RAG-Bots auf existierende unternehmensspezifische LLM ***Foundation-Modellen,*** kann zudem die Entwicklungszeit erheblich verkürzen.[5,8]

Herausforderungen und Risiken

1. **Halluzinationen:** Obwohl RAG-Chatbots insgesamt eine bessere Leistung als herkömmliche Chatbots gezeigt haben, neigen sie zu „Halluzination", d.h. ungenaue oder falsche Informationen zu liefern, die authentisch erscheinen, aber faktisch falsch sind. Dieses Problem tritt bei allen LLM-basierten Systemen auf und wird durch die Verwendung irrelevanter Dokumente in den Antworten noch verschärft.[3,6,11] Eine empirische Studie zu RAG-basierten Rechtsrecherchetools ergab, dass ein großer Teil (zwischen 17 % und 33 %) der Antworten dieser Systeme ungenau oder erfunden war. Oft lieferten sie fehlerhafte rechtliche Erklärungen und erfundene Zitate.[27]

2. **Datenqualität:** Die Effektivität und Zuverlässigkeit von RAG-Systemen hängen stark von der Qualität der Datenbank ab, auf die während der Anfrage zugegriffen wird. Diese wirkt sich direkt auf die Leistung des Chatbots aus.[3] Die für das Training und die Abfrage des RAG-Systems verwendeten externen Wissensquellen sollten so strukturiert sein, dass sie sich für das System leicht identifizieren lassen. Dies zeigt sich insbesondere bei komplexeren Anfragen: Wenn das RAG-System das richtige Dokument nicht finden kann oder die Daten schlecht indexiert sind, gibt der Chatbot möglicherweise eine irrelevante Antwort oder greift auf das Basis-LLM-Modell zurück. Dadurch steigt die Wahrscheinlichkeit, dass das System einen Fehler macht, während die Genauigkeit der Antwort sinkt.[3]

3. **Datenschutz:** Da RAG-Systeme Zugriff auf Datenbanken haben, in denen sich auch sensible Informationen befinden können, ist die Verbreitung privater Daten zu einem großen Problem geworden. Studien zeigen, dass irreführende Texte und Anweisungen in versteckten Inhalten auf Webseiten enthalten sein können. Dadurch werden LLMs gezwungen, diese Anweisungen zu befolgen, wenn sie die Suche über solche „kontaminierten" Webseiten durchführen. Dieser Vorgang wird als „indirekte Prompt-Injektion" bezeichnet. Dadurch können auch die RAG-Systeme personenbezogene Daten abrufen, die versehentlich in das Modelltraining aufgenommen wurden und können sie „übermäßig weitergeben".[6,12] In einem Fall entdeckten Sicherheitsforscher:innen beispielsweise eine Sicherheitslücke im RAG-Assistenten von Microsoft 365 Copilot. Über diese wurden automatisch vertrauliche Dateien (E-Mails, Dokumente usw.) aus den Konten der Benutzer:innen zur Verfügung gestellt.[28]

4. **Übermäßiges Vertrauen in Technologien:** Mehrere Studien haben gezeigt, dass Menschen dazu neigen, KI-Systemen mehr zu vertrauen als Menschen. Ein Phänomen, das als „Automatisierungsbias" bezeichnet wird.[17] Es besteht die Gefahr, dass Menschen die von KI generierten Informationen nicht mehr ausreichend hinterfragen, selbst dann, wenn sie widersprechende Informationen und Wissen über einen konkreten Fall haben.[18]

5. **Transparenz:** Aus technischer Sicht besteht das Problem, dass LLMs aufgrund ihrer Komplexität – selbst für Expert:innen – nicht transparent und nachvollziehbar sind, ein Phänomen das auch als „Black-Box" bezeichnet wird.[33,34] Darüber hinaus erfordert die Nutzung von KI-Tools eine entsprechende Expertise und Schulung, die von den Behörden und Unternehmen, die solche Tools einsetzen wollen, sichergestellt werden müssen (siehe Artikel 4, EU AI-Act).[32]

6. **Urheberrechte:** LLM-Komponenten des RAG-Chatbots werden mit einer Unmenge an Daten trainiert, die oftmals urheberrechtlich geschützt sind. Inwieweit die Nutzung dieser geschützten Inhalte zum KI-Training zulässig ist, wird aktuell kontrovers diskutiert und wird in den kommenden Jahren die Gerichte beschäftigen.[22] Im Jahr 2025 verklagte beispielsweise eine Gruppe von Verlagen (u.a. Forbes und The Guardian) ein KI-Start-up, dessen LLM-Komponente des RAG-Chatbots ihre urheberrechtlich geschützten Artikel kopiert und reproduziert hatte. In seinen Antworten zeigte der Chatbot sogar wörtliche oder zusammengefasste Passagen aus geschützten Nachrichteninhalten an – in einigen Fällen umging er dabei Paywalls.[29]

7. **Verantwortung:** Rechtlich gesehen liegt die Verantwortung bei der Anwendung von RAG-Systemen letztlich bei den Organisationen, die diese Systeme einsetzen und darf nicht auf Nutzer:innen abgewälzt werden. Nach §43 GmbHG greift die Geschäftsführerhaftung bei Sorgfaltspflichtverletzungen.[26] Anders als bei etablierten Technologien, bei denen Verantwortung an qualifizierte Personen delegiert werden kann, existieren für LLMs keine staatlich anerkannten Zertifizierungen oder regulierten Verantwortungsträger:innen. Die Geschäftsführung trägt die Verantwortung somit direkt und kann sie nicht rechtsverbindlich delegieren. Ein Beispiel dafür ist der oben beschriebene Air-Canada-Fall.[11,25]

Empfehlung zum Praxiseinsatz

Um den oben genannten Herausforderungen und Risiken zu begegnen, können folgende Strategien für die öffentliche Verwaltung, Unternehmen und Nutzer:innen empfohlen werden:

Für Behörden und Unternehmen:

- **Ausrichtung an Richtlinien:** Alle Organisationen – sowohl private als auch öffentliche – sind an die Richtlinien gebunden und dürfen mit ihren RAG-Systemen nicht davon abweichen.[20] Insbesondere sollten diese Systeme der Datenschutzgrundverordnung (DSGVO) entsprechen, um die Privatsphäre der Nutzer:innen zu schützen. Die Verarbeitung personenbezogener Daten muss den zentralen Datenschutzanforderungen der Notwendigkeit und Verhältnismäßigkeit entsprechen.[6] Darüber hinaus überwachen Behörden wie das Europäische Büro für Künstliche Intelligenz („AI Office") und nationale Marktüberwachungsbehörden die Umsetzung und Anwendung des *EU AI-Acts*. Artikel 50 des EU AI-Acts schreibt vor, dass Nutzer:innen, die mit einer KI (z.B. einem Chatbot) interagieren, darüber informiert werden müssen, dass sie mit einem KI-System kommunizieren (sofern dies nicht offensichtlich ist).[19]

- **Präzise, hochwertige Daten:** Die Zuverlässigkeit von RAG-Systemen hängt zu einem wesentlichen Teil von der Qualität der Datensätze ab, mit denen LLMs trainiert bzw. im laufenden Betrieb gespeist werden. Die externen Wissensquellen sollen darüber hinaus eindeutig und strukturiert beschriftet werden, um die Genauigkeit zu verbessern. Für die Nutzung von Daten für RAG-Systeme sind weiters folgende Grundsätze einzuhalten:[14,30]

  - **Datenminimierung:** Es wird empfohlen, nur die für die Erfüllung der Aufgabe erforderliche Mindestmenge an personenbezogenen Daten zu verwenden.
  - **Zweckbindung:** Daten, die zur Erbringung von staatlichen Leistungen erhoben werden, sollten nicht willkürlich für andere Zwecke verwendet werden.

- **Menschliche Interventionsmöglichkeit:** Nutzer:innen von RAG-Chatbots sollen stets die Möglichkeit haben, mit menschlichen Mitarbeiter:innen in Kontakt zu treten – insbesondere, wenn die Anfrage komplex, ein persönliches Urteil erforderlich ist oder sie sich über die Richtigkeit der Antwort unsicher sind. Dies verhindert eine ungewollte Automatisierung und unterstützt das Prinzip der „Human in Control"-Aufsicht, wodurch sich das allgemeine Risiko der RAG-Systeme verringern sollte.

- **Technische *Transparenz:*** Ein transparentes RAG-System sollte eindeutig anzeigen, woher die Informationen in der Antwort stammen. Mithilfe der *Erklärbarer KI (XAI)* können Benutzer:innen die zur Generierung von Antworten verwendeten Quellen einsehen. Dadurch soll die Transparenz erhöht und eine kritische Bewertung der Antworten gefördert werden.[31]

- **Use-Case spezifische Gestaltung:** RAG-Systeme müssen genau auf ihren Einsatz zugeschnitten sein. Dafür wird zuerst festgelegt, welche Themen, Nutzergruppen und Dokumentarten erlaubt sind und welche Fragen das System nicht beantworten soll. Dann wird ein passendes Sprachmodell ausgewählt, idealerweise eines, das gut anpassbar ist (z.B. Open-Source). In das System dürfen nur geprüfte und vertrauenswürdige Inhalte aufgenommen werden, damit Fehler und Halluzinationen vermieden werden. Die Qualität der Antworten sollte regelmäßig überprüft werden, zum Beispiel indem man prüft, ob die Informationen relevant und korrekt sind. Außerdem braucht es Schutzmechanismen, damit sensible Daten nicht versehentlich ausgegeben oder falsch verwendet werden.[30,31]

Nutzer:innen (Bürger:innen und Kund:innen):

- **Recht auf *Transparenz:*** Die Benutzer:innen sollen durch einen klar sichtbaren Hinweis erkennen können (durch einen Disclaimer bzw. eine Offenlegung), wenn KI zum Einsatz kommt. Dies dient nicht nur der Transparenz, sondern stärkt auch das Vertrauen in das System und hilft dabei, Antworten kritisch zu bewerten. Dies ist besonders bei RAG-Chatbots wichtig, da ihre Antworten auf einer Kombination aus Sprachmodell und externen Dokumenten beruhen.

- *Datenschutzrechte:* Darüber hinaus haben die Benutzer:innen ein Recht darauf zu erfahren, wie und wann ihre personenbezogenen Daten genutzt und in welche Systeme sie eingespielt werden (Art. 15 DSGVO).[15] Anlaufstelle bei Verdacht auf die Verletzung von Datenschutzrechten ist die österreichische Datenschutzbehörde (DSB).[16]

- **Schaffung und Stärkung des KI-Wissens:** Es ist wichtig, sich über KI-Technologien zu informieren, die richtigen Kompetenzen aufzubauen und die eigenen Rechte zu kennen.[18] In Österreich besteht mit der Rundfunk- und Telekom Regulierungs-GmbH (RTR) eine Informations- und Anlaufstelle für die breite Öffentlichkeit. Die RTR-KI-Servicestelle unterstützt Bürger:innen bei Fragen zur Transparenz, zu Rechten im Umgang mit KI-Systemen oder zur Überprüfung von KI-Entscheidungen.[13]

Wichtige Begriffe

Black-Box-Systeme: Bezeichnet KI-Systeme, deren interne Entscheidungsprozesse für Menschen nicht transparent und nachvollziehbar sind. Das heißt, dass nur Eingaben und Ausgaben beobachtet werden können, ohne zu verstehen, wie die Verarbeitung dazwischen genau abläuft.

Datenschutz in der KI: Die Gesamtheit der Praktiken und Bedenken im Zusammenhang mit der ethischen Erfassung, Speicherung und Nutzung personenbezogener Daten durch Systeme der künstlichen Intelligenz.

Erklärbare KI (XAI): Bezieht sich auf Methoden und Techniken, die es ermöglichen, die Entscheidungen und Ergebnisse von KI-Systemen besser zu verstehen und nachzuvollziehen.

Foundation Modelle: Große, vortrainierte KI-Modelle (z.B. GPT oder BERT), die auf breiten Datenmengen basieren und für vielfältige Aufgaben durch Feinabstimmung angepasst werden können.

Generative KI: Künstliche Intelligenz, die neue Inhalte wie Texte, Bilder, Musik oder Code erzeugen kann, basierend auf erlernten Mustern aus Trainingsdaten.

Gesetz der Europäischen Union über künstliche Intelligenz (EU AI-Act): Eine europäische Verordnung über künstliche Intelligenz (KI), die erste umfassende Verordnung über KI von einer großen Regulierungsbehörde. Sie konzentriert sich insbesondere auf KI-Systeme mit hohem Risiko.

KI-Autonomie: Die Fähigkeit eines KI-Systems, eine Reihe von Zielen unter Berücksichtigung von diversen Unsicherheiten in ihrer Umgebung selbstständig und ohne externe Eingriffe zu erreichen.

KI-Genauigkeit (AI-Accuracy): Bezieht sich auf die Fähigkeit eines KI-Systems, korrekte Vorhersagen oder Entscheidungen zu treffen. Sie ist ein wichtiger Maßstab für ihre Leistung und entscheidend für die Bestimmung ihrer Wirksamkeit und Zuverlässigkeit.

Transparenz: Bedeutet, dass die Funktionsweise, Entscheidungsprozesse und Einsatzbereiche eines KI-Systems nachvollziehbar, erklärbar und offen zugänglich sind – für Entwickler:innen, Nutzer:innen und andere Stakeholder.

Erklärung Stufenmodell des ALAIT Risikoradars

Im ALAIT KI-Risikoradar wird die Beziehung zwischen Anwendungsrisiko und Autonomie eines KI-Systems dargestellt. Die Risikostufen stützen sich auf das EU KI-Gesetz (*EU AI-Act*), insbesondere auf Artikel 6 und Annex III, die sich mit risikoreichen Anwendungsbereichen von KI befassen. Geringere System-Autonomie und Anwendungsrisiken werden durch kältere Farben (blau) und höhere System-Autonomie und Anwendungsrisiken durch wärmere Farben (rot) dargestellt. Der Farbwechsel vermittelt das erhöhte Risiko solcher Entscheidungen. Mithilfe dieser Farbskala lässt sich das Gesamtrisiko erkennen: Violett und dunkelrot – sehr hoch, rot und dunkelorange - hoch, hellorange und Gelbtöne - mittel, Blautöne - geringes Gesamtrisiko. Im Idealfall sollten hohe Anwendungsrisiken und System-Autonomie vermieden oder nur nach sehr sorgfältiger Abwägung eingesetzt werden.

Autonomiegrad des KI-Systems

Stufe 1: Keine Autonomie

KI ist ein passives Werkzeug; Menschen treffen alle Entscheidungen und leiten Maßnahmen ein.

Beispiel: Diagnosesysteme, die medizinische Rohdaten anzeigen oder die Daten analysieren (ohne Empfehlungen!)

Empfohlene Anwendungsfälle: Szenarien mit hohen Risiken oder bei denen ethische Entscheidungen von entscheidender Bedeutung sind (z.B. medizinische Diagnostik, Justizsystem).

Stufe 2: Geringer Autonomiegrad (Human-in-the-Loop)

Die KI gibt Empfehlungen oder Optionen, aber der Benutzer:innen bleibt für die Auswahl und Genehmigung von Maßnahmen verantwortlich.

Beispiel: KI schlägt optimale Routen für die Logistik vor oder Empfehlungssysteme im E-Commerce.

Empfohlene Anwendungsfälle: Aufgaben mittlerer Komplexität mit mäßigen Risiken (z.B. Optimierung der Lieferkette).

Stufe 3: Mittlerer Autonomiegrad (Human-on-the-Loop)

Die KI führt bestimmte Aufgaben autonom aus, wobei Menschen in Ausnahmefällen eingreifen.

Beispiel: KI-gestützte Fertigungsprozesse, bei denen das System Maschinen steuert, aber Nutzende bei Anomalien eingreifen.

Empfohlene Anwendungsfälle: Szenarien, in denen eine kontinuierliche menschliche Beteiligung nicht erforderlich ist, kritische Risiken jedoch eine menschliche Überwachung erfordern (z.B. industrielle Automatisierung, Überwachung von Finanztransaktionen).

Stufe 4: Hoher Autonomiegrad (Human in Control)

Das KI-System arbeitet weitgehend autonom, erlaubt es den Benutzern jedoch, es selbst zu übersteuern, um unerwünschte Ergebnisse zu vermeiden.

Beispiel: Autonome Fahrzeuge

Empfohlene Anwendungsfälle: Umgebungen mit geringem bis mittlerem Risiko (z.B. Logistik, einfaches Verkehrsmanagement).

Stufe 5: Vollständige Autonomie mit minimaler Aufsicht

Das KI-System arbeitet unabhängig und erfordert nur minimale oder gar keine menschliche Intervention. Die Beteiligung des Menschen beschränkt sich auf die langfristige Aufsicht (Audits).

Beispiele: Autonome landwirtschaftliche Maschinen, KI für die Stromnetzverteilung, U-Bahnen, Flughafenbahnen

Empfohlene Anwendungsfälle: Umgebungen mit geringen Sicherheits- oder ethischen Risiken und hoher Zuverlässigkeit des KI-Systems (z.B. sich wiederholende Aufgaben in kontrollierten Umgebungen).

Anwendungsbereich-Risiko

Stufe 1: Minimales Risiko

Das KI-System hat keine Auswirkungen auf die Benutzer:innen oder die Entscheidungsfindung.

Beispiele: Filter, NPCs, Empfehlungsalgorithmen ohne schwerwiegende Folgen (DeepL, andere Übersetzungssysteme)

Kriterien: Keine direkte Auswirkung auf die Hochrisikobereichen des ***EU AI-Acts***.

Stufe 2: Begrenztes Risiko

KI-Systeme, die mit Benutzer:innen interagieren, aber keine Entscheidungen mit hohen Risiken treffen. Das Risiko steigt, wenn es an Transparenz über die Beteiligung von KI mangelt.

Beispiele: Chatbots und KI-generierte Inhalte ohne Offenlegung, einfache Automatisierungsaufgaben.

Kriterien: Bereiche, die nicht in der Liste der „hohen Risiken" des EU AI-Acts enthalten sind.

Stufe 3: Mittleres Risiko

KI-Systeme haben keine besonderen Auswirkungen auf einzelne Personen, aber sie entfalten Wirkung auf kollektiver oder gesellschaftlicher Ebene.

Beispiele: Generative KI wie ChatGPT und andere Systeme, die indirekt die Umgebung beeinflussen können, in der sie eingesetzt werden.

Kriterien: KI-Systeme, die für die öffentliche Nutzung verfügbar sind und das Potenzial haben, bestehende Gepflogenheiten zu beeinflussen und langfristig zu verändern.

Stufe 4: Hohes Risiko

Jeder Algorithmus, der in den laut EU AI-Act „Hochrisikobereichen" angewendet wird oder direkte Auswirkungen auf einzelne Personen hat.

Beispiele: Medizin, Biometrie, kritische Infrastruktur, Bildung und Berufsausbildung, Beschäftigung, Zugang zu Dienstleistungen im öffentlichen Sektor, Strafverfolgung, Migration.

Kriterien: Zugehörigkeit zum „Hochrisikobereich" des EU AI-Acts, nur wenn die Regeln für Transparenz und Datenqualität eingehalten werden.

Stufe 5: Extremes Risiko

Jeder Algorithmus, der in den laut EU AI-Act „Hochrisikobereichen" angewendet wird.

Beispiele: Medizin, Biometrie, kritische Infrastruktur, Bildung und Berufsausbildung, Beschäftigung, Zugang zu Dienstleistungen im öffentlichen Sektor, Strafverfolgung, Migration.

Kriterien: Zugehörigkeit zum „Hochrisikobereich", wenn die Regeln für Transparenz und Datenqualität NICHT eingehalten werden.

Quellen

1 Singh, A. (2025). Agentic RAG Systems for Improving Adaptability and Performance in AI-Driven Information Retrieval. SSRN. https://doi.org/10.2139/ssrn.5188363

2 McKinsey. (2024, October 30). What is RAG (retrieval augmented generation) | McKinsey. https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-retrieval-augmented-generation-rag

3 Jeong, C. (2024). A Study on the Implementation Method of an Agent-Based Advanced RAG System Using Graph. Knowledge Management Research, 25(3), 99–119. https://doi.org/10.15813/kmr.2024.25.3.005

4 Sutter, M. (2025, August 22). Native RAG vs. Agentic RAG: Which Approach Advances Enterprise AI Decision-Making? MarkTechPost. https://www.marktechpost.com/2025/08/22/native-rag-vs-agentic-rag-which-approach-advances-enterprise-ai-decision-making/

5 What is RAG? - Retrieval-Augmented Generation AI Explained - AWS. (n.d.). Amazon Web Services, Inc. Retrieved 19 August 2025, from https://aws.amazon.com/what-is/retrieval-augmented-generation/

6 Retrieval-augmented generation (RAG) | European Data Protection Supervisor. (2025, September 18). https://www.edps.europa.eu/data-protection/technology-monitoring/techsonar/retrieval-augmented-generation-rag

7 Deloitte. (2021). The robots are here: Are you ready?

8 Vohra, D. K. (2024, October 1). How AI and RAG Chatbots Cut Customer Service Costs by Millions. How AI and RAG Chatbots Cut Customer Service Costs by Millions. https://www.nexgencloud.com/blog/case-studies/how-ai-and-rag-chatbots-cut-customer-service-costs-by-millions

9 OneTutor. (n.d.). OneTutor—KI Tutor für Universitäten und Hochschulen. OneTutor. Retrieved 2 October 2025, from https://onetutor.ai

10 Benita, J., Tej, K. V. C., Kumar, E. V., Subbarao, G. V., & Venkatesh, CH. (2024). Implementation of Retrieval-Augmented Generation (RAG) in Chatbot Systems for Enhanced Real-Time Customer Support in E-Commerce. 2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS), 1381–1388. https://doi.org/10.1109/ICACRS62842.2024.10841586

11 Nickodem, K. (2024, May 1). Unpacking the Potential Risks of Generative AI Chatbots on Local Government Websites. Coates' Canons NC Local Government Law. https://canons.sog.unc.edu/2024/05/unpacking-the-potential-risks-of-generative-ai-chatbots-on-local-government-websites/

12 Security Risks with RAG Architectures. (n.d.). IronCore Labs. Retrieved 30 September 2025, from https://ironcorelabs.com/security-risks-rag/

13 KI-Servicestelle der RTR. (n.d.). RTR. Retrieved 16 September 2025, from https://www.rtr.at/rtr/service/ki-servicestelle/ki-servicestelle.de.html

14 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA Relevance), 119 OJ L (2016). http://data.europa.eu/eli/reg/2016/679/oj/eng

15 Art. 15 DSGVO – Auskunftsrecht der betroffenen Person. (n.d.). Datenschutz-Grundverordnung (DSGVO). Retrieved 7 September 2025, from https://dsgvo-gesetz.de/art-15-dsgvo/

16 Datenschutzbehörde, Ö. (n.d.). Österreichische Datenschutzbehörde. Österreichische Datenschutzbehörde. Retrieved 5 September 2025, from https://dsb.gv.at/

17 Wihlborg, E., Larsson, H., & Hedström, K. (2016). 'The Computer Says No!' – A Case Study on Automated Decision-Making in Public Authorities. 2016 49th Hawaii International Conference on System Sciences (HICSS), 2903–2912. https://doi.org/10.1109/HICSS.2016.364

18 Thapa, B. (2019). Predictive Analytics and AI in Governance: Data-driven government in a free society. The European Liberal Forum. https://liberalforum.eu/publication/predictive-analytics-and-ai-in-governance-data-driven-government-in-a-free-society/

19 Article 50: Transparency Obligations for Providers and Deployers of Certain AI Systems | EU Artificial Intelligence Act. (n.d.). Retrieved 14 October 2025, from https://artificialintelligenceact.eu/article/50/

20 Richtlinie—2024/2853—DE - EUR-Lex. (n.d.). Retrieved 15 October 2025, from https://eur-lex.europa.eu/eli/dir/2024/2853/oj/deu

21 Digital government. (n.d.). OECD. Retrieved 15 October 2025, from https://www.oecd.org/en/topics/digital-government.html

22 Künstliche Intelligenz: Eine Herausforderung für das Urheberrecht. (n.d.). Retrieved 15 October 2025, from https://www.wu.ac.at/forschung/forschungsportal/news/archiv-news/detail/kuenstliche-intelligenz-eine-herausforderung-fuer-das-urheberrecht

23 RAG-Chatbot: Technische Dokumentation. (n.d.). RTR. Retrieved 17 October 2025, from https://www.rtr.at/rtr/service/ki-servicestelle/chat/technik.de.html

24 NYC's AI chatbot was caught telling businesses to break the law. The city isn't taking it down. (2024, April 3). AP News. Retrieved 21 October 2025, from https://apnews.com/article/new-york-city-chatbot-misinformation-6ebc71db5b770b9969c906a7ee4fae21

25 Cecco, L. (2024, February 16). Air Canada ordered to pay customer who was misled by airline's chatbot. The Guardian. https://www.theguardian.com/world/2024/feb/16/air-canada-chatbot-lawsuit

26 §43 GmbHG – Einzelnorm. (n.d.). Retrieved 31 October 2025, from https://www.gesetze-im-internet.de/gmbhg/__43.htm

27 Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. Journal of Empirical Legal Studies, 22(2), 216–242. https://doi.org/10.1111/jels.12413

28 NewsBites Volume XXVII – Issue 45, June 13, 2025. (2025, June 13). SANS Institute. https://www.sans.org/newsletters/newsbites/xxvii-45

29 Tsai, L., & Tsai, P.-J. (2025, May 29). Generative AI Copyright Lawsuit: RAG Technology Once Again in Focus as News Publishers Sue Cohere. Lexology. https://www.lexology.com/library/detail.aspx?g=7a1f5a5b-0274-4fdc-8135-7bb5a80671f3

30 Kelbert, T. H., Dr Julien Siebert, Patricia. (2024, May 13). Retrieval Augmented Generation (RAG): Chat mit eigenen Daten. Fraunhofer IESE. Retrieved 5 November 2025, from https://www.iese.fraunhofer.de/blog/retrieval-augmented-generation-rag/

31 Alrabie, L., & Andolina, S. (2025). Towards Human-Centered RAG: A Study of AI-Supported Testing Practices in Italian Public Administration. Proceedings of the 16th Biannual Conference of the Italian SIGCHI Chapter, 1–6. https://doi.org/10.1145/3750069.3750103

32 Article 4: AI literacy | EU Artificial Intelligence Act. (n.d.). Retrieved 5 September 2025, from https://artificialintelligenceact.eu/article/4/

33 Thapa, B. (2019). Predictive Analytics and AI in Governance: Data-driven government in a free society. The European Liberal Forum. https://liberalforum.eu/publication/predictive-analytics-and-ai-in-governance-datadriven-government-in-a-free-society/

34 Santiso, C. (n.d.). Public Governance in the Age of Artificial Intelligence. Retrieved 1 July 2025, from https://www.chandlerinstitute.org/governancematters/public-governance-in-the-age-of-artificial-intelligence

Projekt ALAIT

Das Austrian Lab for AI Trust (ALAIT) ist ein vom österreichischen Bundesministerium für Innovation, Mobilität und Infrastruktur (BMIMI) initiiertes Forschungs- und Entwicklungs-Projekt zur Schaffung von Vertrauen durch Wissen im Bereich Künstliche Intelligenz (KI). Das Projekt ALAIT zielt darauf ab, Interessierte und wichtige gesellschaftliche Gruppen zu befähigen, KI-Technologien verantwortungsvoll zu nutzen und ethische sowie qualitativ hochwertige Standards für den Einsatz von AI zu etablieren.

Das Projekt wird von **winnovation** geleitet (Gertraud Leimüller und Lena Müller-Kress) und im Konsortium mit **leiwand.ai** (Rania Wazir und Silvia Wasserbacher-Schwarzer), **TU Wien** (Sabine Köszegi und Ilya Faynleyb) und **Austria Presse Agentur – APA** (Verena Krawarik und Sophia Marecek) umgesetzt.

Die ALAIT-Dossiers sind auf der Projekthomepage abrufbar: https://science.apa.at/project/alait/

Die Inhalte des Dossiers entsprechen dem aktuellen Stand der Technik und wurden sorgfältig nach wissenschaftlichen Kriterien erstellt. Sie dienen jedoch nicht als rechtsverbindliche Auskunft oder Beratung.

Impressum

Medieninhaberin und Herausgeberin:
winnovation consulting gmbh
Linke Wienzeile 42/1, Top 5
1060 Vienna

Danksagung:

Wir danken folgenden Expert:innen für ihr hilfreiches Feedback zu Vorversionen dieses Dossiers: Thomas Schreiber, Brigitte Krenn, Martin Berninger.