SEQIS QualityNews

Analysis. Development. Test. Management

Ausgabe H2/2025

AI Act & Local AI - Wege in eine verantwortungsvolle Zukunft

Der unsichtbare Kollege: Künstliche Intelligenz — Seite 4

Tschüss Cloud, Hallo Souveränität: die Technologie hinter Local AI — Seite 14

razzfazz.ai - local matters — Seite 19

Bild generiert von Gemini (Google AI)

Analysis. Development. Test. Management. Better Results.

Lesen Sie in dieser Ausgabe:

Editorial.....3

Der unsichtbare Kollege: Künstliche Intelligenz.....4
Tanja Huber

Artificial Intelligence: Vom Hype zur echten Wertschöpfung.....8
Alexander Weichselberger

Local AI: Vorteile / Nachteile.....12
Ahmed Soliman

Tschüss Cloud, Hallo Souveränität: die Technologie hinter Local AI.....14
Martin Brandhuber

razzfazz.ai - local matters.....19
Alexander Vukovic

Der AI Act im Jahr 2025 - Was bisher geschah.....28
Alexander Lewisch

KI-Souveränität im Fokus: Wie Local AI die Datenhoheit sichert und den EU AI Act erfüllt.....36
Martin Wildbacher

Die Konvergenz von KI-Ethik und dem AI Act: Von der Abstraktion zur Anwendung.....42
Alexander Lewisch

KI im Unternehmen: Ein Praxis-Leitfaden zu Datenschutz, AI Act und Compliance.....47
Emre Kurtulush

Denken trotz KI: Warum Automatisierung den Verstand nicht ersetzen darf.....60
Amine Boutahar

Referenzstory: Otto Bock Healthcare Products GmbH.....62
mit Manius Schinkel

Lokales AI-Hosting auf dem Mac: Ein praktischer Leitfaden für Coder.....64
Daniel Kleissl

5 Fragen an die KI-Experten: Wie künstliche Intelligenz die Unternehmenswelt revolutioniert.....70
Sandra Benseler

Über SEQIS QualityNews:
Dieses Magazin richtet sich an Gleichgesinnte aus den Bereichen IT Analyse, Development, Softwaretest und Projektmanagement im IT Umfeld. Die SEQIS Experten berichten über ihre Erfahrungen zu aktuellen Themen in der Branche. Die Leser des Magazins gestalten die Ausgaben mit: Schreiben Sie uns Ihre Meinung im SEQIS Blog (www.SEQIS.com/de/blog-index) oder als Leserbrief.
Wenn Sie dieses Magazin abbestellen möchten, senden Sie bitte ein Mail an marketing@SEQIS.com.

Impressum:
Information und Offenlegung gem. §5 E-Commerce-Gesetz und §25 Mediengesetz

Herausgeber: SEQIS GmbH, Wienerbergstraße 3-5, A.1100 Wien
info@SEQIS.com, www.SEQIS.com
Gericht: Handelsgericht Wien
Firmenbuchnummer: 204918a
Umsatzsteuer-ID: ATU51140607
Geschäftsführung: Mag. (FH) Alexander Vukovic, Mag. (FH) Alexander Weichselberger, DI Reinhard Salomon

Druck: druck.at Druck- und Handels-gesellschaft mbH, 2544 Leobersdorf
Erscheinungsweise: 2x pro Jahr

DI Reinhard Salomon | Mag. (FH) Alexander Vukovic | Mag. (FH) Alexander Weichselberger

Editorial

Sehr geehrte Leserin,
sehr geehrter Leser,

wir freuen uns, Ihnen die Ausgabe für das zweite Halbjahr 2025 zu präsentieren.

Vielen Dank für das positive Feedback zu unserer letzten Ausgabe zu dem Themenbereich „Nachhaltigkeit in der IT“. Wir hoffen, wir konnten Ihnen damit interessanten Content zur Verfügung stellen und Sie stellenweise auch gut unterhalten. Über weitere Anregungen, Themenwünsche und Feedback Ihrerseits freuen wir uns.

Auch in dieser Ausgabe finden Sie neben den branchenbezogenen Artikeln auch nicht-technische Bereiche:

In dieser Ausgabe dreht sich alles um das Thema „AI Act und Local AI“. Wie immer bieten wir eine Auswahl von Fachartikeln mit diesem Schwerpunkt, die wir aus unserem blog.seqis.com für Sie zusammengestellt haben.

Auf den folgenden Seiten geben Ihnen unsere Experten einen Einblick in die vielseitigen Aspekte zum Thema.

Wir wünschen Ihnen viel Lesevergnügen mit der aktuellen Ausgabe der SEQIS QualityNews!

Ihre SEQIS Geschäftsleitung

Der unsichtbare Kollege: Künstliche Intelligenz

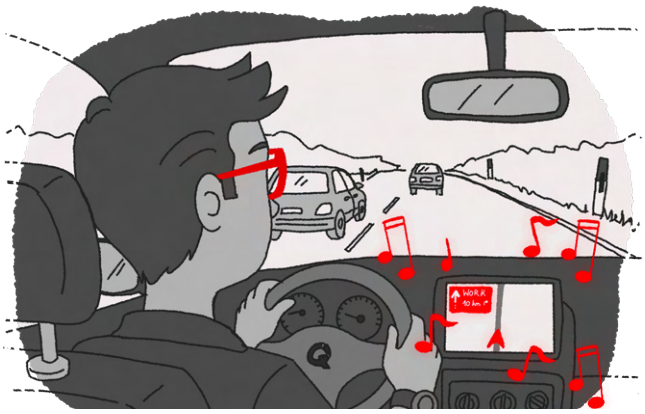
von Tanja Huber

Sind Sie sich überhaupt bewusst, wie oft einem Durchschnittsverbraucher heutzutage über einen normalen (Arbeits-)Tag Künstliche Intelligenz (KI) begegnet? Es ist weit mehr, als man ursprünglich annehmen würde? Sehen wir uns doch mal ein Beispiel an!

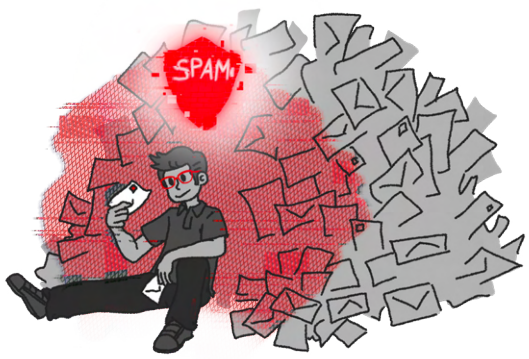
In unserem Beispiel ist das hier unser Otto Normalverbraucher: Im Alter zwischen 30 und 40 Jahren, da diese Altersgruppe Technologie-kompetenter ist und Apps, Streaming als auch Smart Home verwendet. Aber bereits wirtschaftlich etabliert, sodass unser Otto sich diese Geräte und Services auch leisten kann. Wir nehmen an, dass er sowohl Smartphone- als auch Smartwatch-Besitzer ist.
Unser Beispiel-Kollege ist ein Büroangestellter (ob Marketing, Vertrieb, IT, usw. ist für unser Beispiel mal nicht von Bedeutung), der täglich von seinem Wohnort zu seinem Arbeitsplatz pendelt. Für unser Beispiel verfolgen wir Otto an einem normalen Arbeitstag, aber er freut sich bereits auf seinen nächsten Urlaub.
Somit haben wir unseren generischen Charakter, mit dem wir durch unseren Beispiel-Tag gehen werden.

06:00 Uhr
Otto wird sanft aufgeweckt - basierend auf Daten seiner KI-Schlafanalyse, die seinen Zyklus gelernt und basierend darauf eine Voraussage für die ideale Schlafphase zum Wecken erstellt hat.
Seine Smartwatch trackt seine Schlafphasen und sammelt diese Daten.

07:00 Uhr
Otto sitzt am Frühstückstisch und während er isst, fragt er seinen Sprachassistenten der Wahl, wie das Wetter sich heute und im Lauf der Woche entwickeln wird - er freut sich schließlich schon auf seinen bevorstehenden Urlaub und wünscht sich dafür gutes Wetter.
Die KI, die hierbei zum Einsatz kommt, verbessert nicht nur die Rohdaten und erkennt komplexe Muster (u.A. auf der Grundlage historischer Daten), sondern gibt die Ergebnisse auch lokal präzise aus. Dadurch wird die Wettervorhersage erheblich beschleunigt und vereinfacht.



07:30 Uhr
Auf dem Weg zur Arbeit nutzt Otto einen KI-Routing Algorithmus, um die schnellste Route zu seinem Ziel basierend auf Echtzeitdaten zu erhalten. Während er fährt, lässt er Lieder basierend auf KI-Empfehlungen im Musikstreaming-Dienst seiner Wahl laufen, die ihm den "passenden" Song vorspielen.
Zwischendrin lässt Otto sich durch seinen Sprachassistenten seine heutigen Termine aufzählen, damit er sich bereits geistig darauf vorbereiten kann.



09:00 Uhr
Während Otto konzentriert arbeitet, bewahrt ihn sein KI-Spam-Filter vor 99 % der unerwünschten E-Mails, die sonst seinen Posteingang überfluten würden.
Beim Schreiben seiner Mails unterstützt ihn die KI-Textvorhersage, indem sie die nächsten Worte vorschlägt. Da Otto mit seinen Formulierungen aber immer noch nicht ganz zufrieden ist, lässt er die KI seinen Text umformulieren, bis er mit dem Ergebnis zufrieden ist.

12:30 Uhr
Es ist Zeit für Ottos Mittagspause. Heute geht er wieder zu seinem Lieblingsbäcker, wo er erneut mit Künstlicher Intelligenz in Berührung kommt. Der Bäcker nutzt einen Algorithmus, der basierend auf Wetterdaten, Wochentag und historischen Verkaufszahlen die Tagesproduktion optimiert und somit Abfall minimiert. Er zahlt sein Essen kontaktlos mit seiner Karte, wobei im Hintergrund eine KI-gestützte Betrugserkennung durchgeführt wird. Diese analysiert das normale Kaufmuster und entscheidet, ob eine Transaktion ungewöhnlich ist.

15:00 Uhr
Otto nimmt sich ein wenig Zeit für eine kleine Online-Recherche für sein Traumurlaubsziel. Seine Suchergebnisse sind basierend auf früheren Suchanfragen und seinen Daten (z.B. Alter, Standort, besuchte Webseiten) angepasst. Beispielsweise hatte er in früheren Suchanfragen teurere Hotels bevorzugt - somit werden ihm diese nun zuerst vorgeschlagen.
Beim Öffnen des ersten Suchergebnisses stößt Otto auf KI-gesteuerte, personalisierte Werbung. Auch die angezeigten Preise für die Reisen wurden in Echtzeit analysiert, um ihm die beste und günstigste Option anzubieten

18:00 Uhr
Otto schließt seine Arbeit ab und beschließt, vor der Heimfahrt noch schnell ein paar Lebensmittel einzukaufen und sucht seinen nächsten Supermarkt auf. Dort finden KI-gestützte Kassensysteme und Inventur-KIs Einsatz, um die Regale stets optimal gefüllt zu halten und alle Waren verfügbar zu machen.

20:00 Uhr
Um seinen Abend abzurunden, will Otto einen Film bei seinem Lieblings-Streaming-Dienst ansehen. Die KI-Empfehlungs-Engine schlägt ihm direkt eine Liste mit auf seinen bisherigen Sehgewohnheiten basierenden Filmauswahl vor. Am Ende wählt er einen Film aus, der an seinem Urlaubsziel spielt, um sich so auf seine Reise einzustimmen.

Man sieht: An mehreren Stellen von Ottos Alltag gibt es KI, die man im ersten Moment nicht erwarten würde oder auf die Otto gar keinen Einfluss hat. Vielerorts macht künstliche Intelligenz seinen Alltag einfacher, ohne dass Otto das überhaupt wahrnimmt. Das zeigt, wie tief KI bereits in unserem täglichen Leben verwurzelt ist und gar nicht mehr wegzudenken ist.

Neben Otto gibt es viele Mitarbeiter und Kollegen, die sich bewusst für den Einsatz von KI entscheiden und mit vollem Bewusstsein diverse Chatbots nutzen (und nutzen wollen), um ihre Arbeit zu erleichtern. Viele Unternehmen zögern allerdings noch mit einer (geregelten) Einführung von KI, wodurch viele der Kollegen privat und ohne Wissen des Unternehmens KI nutzen, die potentiell eine große Sicherheitslücke für das Unternehmen darstellen könnte. Denn der einzelne Mitarbeiter ist sich der Auswirkungen der Nutzung von KI auf den Datenschutz und die IT-Sicherheit eines Unternehmens oft nicht bewusst und denkt auch nicht darüber nach - das ist auch nicht seine Aufgabe. Diese Verantwortung liegt beim Unternehmen selbst.

Allein beim Beispiel von Otto kann man bereits einige potentielle Sicherheitslücken finden, so unwahrscheinlich sie vielleicht sein mögen. Sehen wir uns schnell 2 Fälle an:

- Seine Smartwatch sammelt eine Unmenge Daten über ihn, die bei einem Hack (den man nie vollkommen ausschließen kann) Angreifern detaillierte Einblicke in seine Anwesenheitszeiten, Tagesstruktur und seinen Standort ermöglichen. Diese Informationen können für Social-Engineering-Angriffe gegen ihn oder sein Unternehmen missbraucht werden.

- Während der Fahrt greift Otto über den Sprachassistenten seines privaten Smartphones auf seinen geschäftlichen Kalender zu. Da er während der Fahrt nicht lesen kann, lässt er sich alle Termine vorlesen. Termine könnten dabei sensible Namen wie „Strategiemeeting mit Kunde X" haben. Sollte die Aufzeichnung oder das Protokoll des Sprachassistenten gehackt werden, würden vertrauliche Geschäftsinformationen über einen nicht autorisierten, privaten Kanal offengelegt.

Potenzial allein garantiert nicht, dass ein Risiko zur Realität wird. Es braucht allerdings nur einmal etwas schiefgehen, um weitreichende Konsequenzen mit sich ziehen. Ein bewusster und geregelter Umgang mit Künstlicher Intelligenz ist somit stets vorzuziehen, um mögliche negative Auswirkungen zu minimieren und die Vorteile dieser Technologie optimal zu nutzen. Es ist sehr zu empfehlen, proaktiv mögliche Risiken zu identifizieren, zu bewerten und Strategien zu entwickeln, damit umzugehen.

Doch wie schafft man das?
Zunächst einmal: Von einem Verbot für die Nutzung von KI sollte abgesehen werden, denn diese führen oft nur zu einer heimlichen und unkontrollierten Verwendung. Verbote führen nur zur Nutzung des Verbotenen in ungeregelten Ausmaßen. Die Mitarbeiter, welche KI nutzen, wollen zudem die Vorteile von KI nutzen. Durch eine Entlastung von repetitiven Aufgaben durch KI gewinnen sie wertvolle Zeit für anspruchsvollere und kreativere Tätigkeiten. Und: Unternehmen, die KI meiden, schließen sich von Effizienzgewinnen aus und riskieren, von agileren Wettbewerbern überholt zu werden.

Doch was tut man da jetzt nun?
Das Unternehmen muss klare Regeln und Richtlinien definieren, an denen sich die Mitarbeiter orientieren können. Dadurch wird nicht nur ein transparentes Arbeitsumfeld geschaffen, sondern auch die Effizienz und Konsistenz der Arbeitsabläufe gefördert. Wenn Mitarbeiter wissen, welche Verhaltensweisen im Unternehmen gewünscht sind und welche Produkte genutzt werden können, können unerwünschte Folgen und Fehlverhalten verringert werden.

Beispiele für eine unternehmensweite Strategie für KI könnte folgendes sein:

- Die Einführung eines "KI-Führerscheins" bzw. einer Schulung oder Zertifizierung, der den Mitarbeitern das nötige Wissen und die Fähigkeiten vermitteln soll, KI sicher und effizient im Arbeitsalltag zu nutzen.

- Das Etablieren einer offenen Kommunikationskultur, in der Mitarbeiter ermutigt werden, Fragen zu stellen und ihre selbst entdeckten Tools der IT-Abteilung für eine Abklärung zu melden.

- Das Definieren einer Whitelist an genehmigten KI-Tools (z.B. eine gesicherte interne Unternehmens-KI) und ein explizites Verbot der Nutzung ungesicherter, öffentlicher Modelle.

- Die Nutzung eines Vier-Augen-Prinzips bevor etwas von einer KI generierter Inhalt verwendet oder gar veröffentlicht wird.

- Das Einrichten eines interdisziplinären Teams (IT, Recht, Fachbereiche), das Tools prüft, freigibt und sicher bereitstellt.

Künstliche Intelligenz ist aus unserem Alltag nicht mehr wegzudenken, wie Ottos Beispiel zeigt. Sie ist bereits fest in unser Leben und unseren täglichen Abläufen integriert. Anstatt sich ihr zu widersetzen, sollte KI strategisch genutzt werden. Die Frage ist nicht mehr, **ob** KI eingesetzt wird, sondern **wie**. Es ist entscheidend, diese Frage nicht dem Zufall zu überlassen, sondern sich aktiv mit den Möglichkeiten und Herausforderungen auseinanderzusetzen, um ihr volles Potenzial auszuschöpfen. Handeln Sie jetzt und überdenken Sie Ihre Strategien - denn KI ist längst im Haus.

Ein besonderes Highlight:

Alle Abbildungen dieses Beitrags wurden mit viel Talent und Hingabe von der Autorin selbst entworfen und gezeichnet.

Tanja Huber ist Consultant bei SEQIS.

Zusätzlich zu ihrer Erfahrung in den Bereichen Testfallerstellung, Testdurchführung und Testunterstützung ist sie auch in den Bereichen Projekt- und Qualitätsmanagement tätig.

Beim Herangehen an neue Aufgaben ist ihr eine gewissenhafte Vorgehensweise und kreative Aufgabenlösung wichtig. Ebenso legt sie größten Wert auf hochwertige Qualitätssicherung.

Artificial Intelligence: Vom Hype zur echten Wertschöpfung

von Alexander Weichselberger

Wie wir KI im Unternehmen sinnvoll verankern

Kennen Sie das „Bullshit-Bingo"? Früher warteten wir in Meetings darauf, Begriffe wie „Disruption" oder „VUCA" abhaken zu können. Heute würde auf fast jedem Zettel ganz oben „KI" oder „Artificial Intelligence" stehen. Doch jenseits der Schlagworte stellt sich für uns Digitalisierer, Analysten und Projektleiter die entscheidende Frage: Wie vermeiden wir, dass KI zur neuen „Shelfware" wird – Software, die wir begeistert kaufen, aber am Ende im Regal verstaubt, weil der echte Nutzen fehlt?

Die Technologie ist da. Aber zwischen dem bloßen Herumspielen und einer wertschöpfenden Integration in kritische Geschäftsprozesse liegt ein weiter Weg. Es geht nicht mehr um das Ob, sondern um das Wie. Wir müssen KI professionell, sicher und zielgerichtet einsetzen, ähnlich wie wir es bei der Testautomatisierung oder RPA gelernt haben: Fokus auf den ROI und weg von reiner Technik-Verliebtheit.

Strategie vor Technologie: Das „Warum" definieren

Der häufigste Fehler ist der „Solutionism" – man hat eine Lösung (AI) und sucht nun krampfhaft nach einem Problem. Eine erfolgreiche Implementierung beginnt jedoch immer mit der Unternehmensstrategie. Bevor die erste Zeile Code geschrieben wird, müssen wir definieren, was wir erreichen wollen. Ziele müssen messbar sein und direkt auf die Vision einzahlen. Für uns bedeutet das, Metriken jenseits von „Modellgenauigkeit" zu finden. Wir müssen uns fragen: Senkt der Einsatz die Durchlaufzeiten? Reduzieren wir die Fehlerquote („First Time Right")? Oder verbessern wir den Customer Effort Score im Support? Nur wenn der Return on Investment (ROI) stimmt, hat das Projekt eine Zukunft.

Der Rahmen: Compliance und Standards als Qualitätsmerkmal

KI-Einführung ist primär Change Management. Es reicht nicht, Data Scientists einzustellen; wir müssen eine breite KI-Kompetenz im Unternehmen aufbauen. Fachabteilungen müssen verstehen, was KI kann – und was nicht. Hier fungieren Analysten als unverzichtbare Übersetzer zwischen Business und Tech, um „mündige Keyuser" zu etablieren, die Technologie als Werkzeug begreifen. Dabei ist der kommende EU AI Act kein Hindernis, sondern ein Qualitätsmerkmal. Er zwingt uns dazu, Risiken frühzeitig zu klassifizieren und Transparenz zu schaffen. Ähnlich wie wir bei Remote Work gelernt haben, dass „Security First" gilt, müssen wir bei KI sicherstellen, dass wir keine „Black Box" betreiben.

Flankierend dazu liefert die ISO/IEC 42001:2023 das notwendige Werkzeug. Ich halte diese Norm für einen enorm starken „Supportive": Als erster globaler Standard für AI Management Systems (AIMS) übersetzt sie die regulatorischen Anforderungen in greifbare Prozesse. Für uns Quality Engineers ist die ISO 42001 das Rückgrat, um Compliance operativ umzusetzen. Sie gibt uns Struktur, ähnlich wie wir es von Qualitätsmanagement-Normen kennen, und hilft uns, KI nicht nur gesetzeskonform, sondern auch vertrauenswürdig und risikominimiert zu betreiben.

Der Kurs "EU AI Act in der Praxis" bietet einen strukturierten Fahrplan, um die neuen regulatorischen Anforderungen nicht nur zu verstehen, sondern als strategische Chance für das Unternehmen zu nutzen.

Sie erhalten in drei praxisnahen Modulen:

- Von den technischen Grundlagen
- Über die rechtlichen Details
- Bis zur konkreten Implementierung

das notwendige Rüstzeug inklusive Checklisten, um KI-Systeme sicher und gesetzeskonform zu betreiben.

Das Training steht sowohl als offenes Format (Online/Onsite) für Einzelpersonen als auch als maßgeschneiderte Inhouse-Lösung für Teams zur Verfügung und schließt mit einem qualifizierten Zertifikat als Kompetenznachweis ab.

Weitere Details siehe https://seqis.ai/de/

Die fünf Säulen einer qualitätsgesicherten KI-Einführung

Wer sich ernsthaft mit Softwarequalität auseinandersetzt, weiß: Ein System ist nur so gut wie sein schwächstes Glied. Bei KI gelten fünf unverhandelbare Prinzipien.

Abbildung 1: Die 5 Säulen einer qualitätsgesicherten KI-Einführung
Quelle: Bild generiert von Gemini (Google AI) - Texte von SEQIS GmbH

- Erstens müssen wir realistische Erwartungen managen. KI ist Mathematik, keine Magie. Als IT-Verantwortliche müssen wir die Erwartungshaltung im Management dämpfen: KI halluziniert, macht Fehler und versteht keinen Kontext, den sie nicht gelernt hat.
- Zweitens gehört der Mensch in den Mittelpunkt (Human-in-the-loop). Die Angst, ersetzt zu werden, ist real. Unsere Botschaft muss lauten: KI soll assistieren, nicht eliminieren. Wir nutzen Automatisierung, um repetitive Aufgaben zu erledigen, damit Zeit für wertschöpfende Tätigkeiten bleibt.
- Drittens steht und fällt alles mit der Datenqualität. Das alte Prinzip „Garbage In, Garbage Out" gilt hier mehr denn je. Ein Modell, das mit bias-behafteten Daten trainiert wird, liefert schädliche Ergebnisse – hier müssen wir Vorarbeiten optimieren und Daten strukturieren, bevor wir automatisieren.
- Viertens brauchen wir eine kontinuierliche Verbesserung. Software ist deterministisch, KI ist dynamisch. Ein Modell, das heute gut ist, kann morgen nutzlos sein. Wir brauchen ein ständiges Monitoring, ähnlich wie wir Produktionssysteme überwachen.

Und fünftens ist Sicherheit nicht verhandelbar. KI-Modelle sind neue Angriffsvektoren. Wir müssen Security by Design von Anfang an mitdenken, um unsere Systeme gegen Manipulationen abzuhärten.

Ein Blick in die Praxis: Wo die Reise hingeht

Um den Appetit auf die Umsetzung zu wecken, lohnt sich ein Blick auf die konkreten Chancen, die sich uns bieten. In der Fertigungsindustrie wird die Effizienz greifbar: Prädiktive Wartung analysiert Sensordaten, um Ausfälle vorherzusagen, bevor sie passieren, während Bilderkennungssysteme in der Qualitätskontrolle Defekte schneller und genauer finden als das menschliche Auge.

Im Finanzsektor und Einzelhandel geht es um Präzision und Personalisierung. Algorithmen erkennen Betrugsversuche in Echtzeit oder bewerten Kreditrisiken differenzierter als starre Scorecards. Gleichzeitig ermöglicht KI im Marketing eine Hyper-Personalisierung, die dem Kunden genau das empfiehlt, was er braucht, statt Gießkannen-Werbung zu betreiben – ähnlich wie wir im Service-Bereich durch Automatisierung schnellere Reaktionszeiten für den Kunden schaffen.

Abbildung 2: Quelle: Bild generiert von Gemini (Google AI)

Besonders spannend ist der Einsatz im Gesundheitswesen. Hier fungiert KI als „zweites Augenpaar" für Ärzte, analysiert Röntgenbilder zur Diagnoseunterstützung und beschleunigt die Entwicklung neuer Medikamente massiv. Und nicht zuletzt optimiert KI in der Logistik Lieferketten und Routen so effizient, dass Kosten gesenkt und Lieferzeiten verbessert werden.

Fazit: Start Small, Scale Fast

Die Einführung von KI ist kein Sprint, sondern ein Marathon. Fangen wir mit kleinen, überschaubaren Pilotprojekten an, bei denen die Datenlage gut ist und der Nutzen schnell sichtbar wird – dort, wo wir die „Hauptstraßen der Wertschöpfung" finden. Lernen wir daraus, passen wir unsere Strategie an und skalieren dann. Die Werkzeuge liegen bereit – es liegt an uns, sie klug und qualitätsgesichert zu nutzen.

Alexander Weichselberger ist Managing Partner.

Er hat seine Einsatzschwerpunkte in den Bereichen Systemanalyse, Softwaretest, Koordination und Management von exponierten Großprojekten und kann auf jahrelange Erfahrung zurückblicken.

Dieses Wissen gibt er gerne in Form von Coachings, Methodentrainings und Fachvorträgen weiter.


IT PingPong
AI-Specials
Jetzt anhören!
Überall da, wo
es Podcasts gibt.
IT PingPong
DeepSeek & LLMs:
Was steckt
dahinter?
IT PingPong
Artificial
Intelligence
im Fokus
IT PingPong
AI-Agenten:
Digitale Helfer
im Alltag
IT PingPong
AI Act-Klartext
Was kommt auf
uns zu?
IT PingPong
AI 101
for everyone
Zu den Folgen:

Local AI: Vorteile / Nachteile

von Ahmed Soliman

Künstliche Intelligenz (KI) ist längst kein Zukunftsthema mehr, sondern Teil unseres Arbeitsalltags. Meist nutzen wir KI in der Cloud – also über Dienste großer Anbieter. Doch immer häufiger wird die Frage gestellt: Sollte KI nicht auch lokal betrieben werden? Diese sogenannte Local AI bringt interessante Vorteile, aber auch klare Herausforderungen mit sich[1].

Was bedeutet Local AI?

Unter Local AI versteht man den Betrieb von KI-Modellen direkt auf der eigenen Hardware – etwa auf Firmenservern oder leistungsstarken Workstations – statt in einer externen Cloud.[2]
Das können kleine Sprachmodelle sein, die auf einem Laptop laufen, oder komplexere Lösungen auf Servern mit starker GPU-Unterstützung.

Vorteile von Local AI

Abbildung 1: Quelle: Bild generiert von OpenAI

1. **Datensicherheit & Privatsphäre**
   Sensible Daten verlassen das eigene Unternehmen / Einflusssphäre nicht. Damit entfällt das Risiko, dass Informationen durch Cloud-Anbieter gespeichert oder ausgewertet werden könnten.[3]

2. **Unabhängigkeit von externen Hosting-Partnern**
   Keine Abhängigkeit von Preismodellen und -anpassungen, Nutzungsgrenzen oder möglichen Ausfällen externer Services.[1]

3. **Offline-Funktionalität**
   Local AI kann auch ohne Internetverbindung genutzt werden – z. B. in abgeschotteten Netzwerken oder bei mobilen Einsätzen.

4. **Schnelligkeit**
   Da keine Übertragung ins Internet notwendig ist, entfallen Netzwerklatenzen.[2]

5. **Anpassbarkeit**
   Modelle lassen sich speziell auf die eigenen Daten trainieren (Fine-Tuning). Damit können individuelle, stets gleiche Lösungen entstehen, die ein Cloud-Service nicht abdeckt.

Nachteile von Local AI

1. **Hoher Hardware-Bedarf**
   Moderne Sprachmodelle benötigen GPUs, viel RAM und Speicher. Diese Anschaffung ist teuer.[2]

2. **Wartung & Betrieb**
   Updates, Sicherheitspatches und Modellpflege liegen in der Verantwortung des Unternehmens.[1]

3. **Begrenzte Skalierbarkeit**
   Während Cloud-Dienste flexibel Ressourcen bereitstellen, stößt lokale Hardware schnell an ihre Grenzen, wenn viele Nutzer gleichzeitig zugreifen.

4. **Energieverbrauch**
   Eigenbetrieb bedeutet zusätzlichen Strombedarf und damit höhere Betriebskosten - wobei man diese den externen Kosten gegenüberstellen müsste.[3]

5. **Funktionsumfang**
   Viele Cloud-Dienste bieten Features (z. B. multimodale Modelle, automatische Skalierung), die lokal nur schwer verfügbar sind.

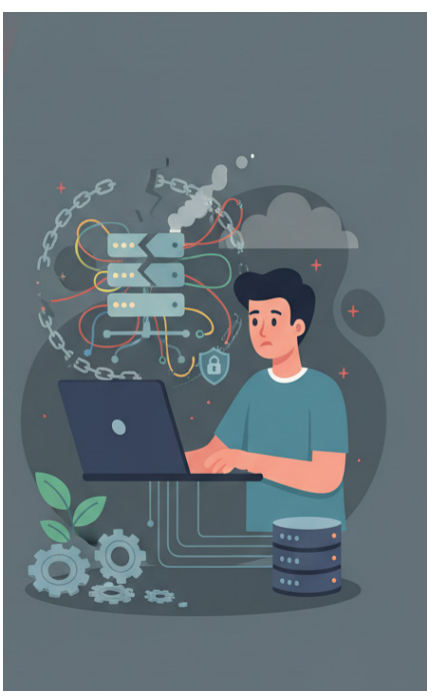


Abbildung 1: Quelle: Bild generiert von Gemini (Google AI)

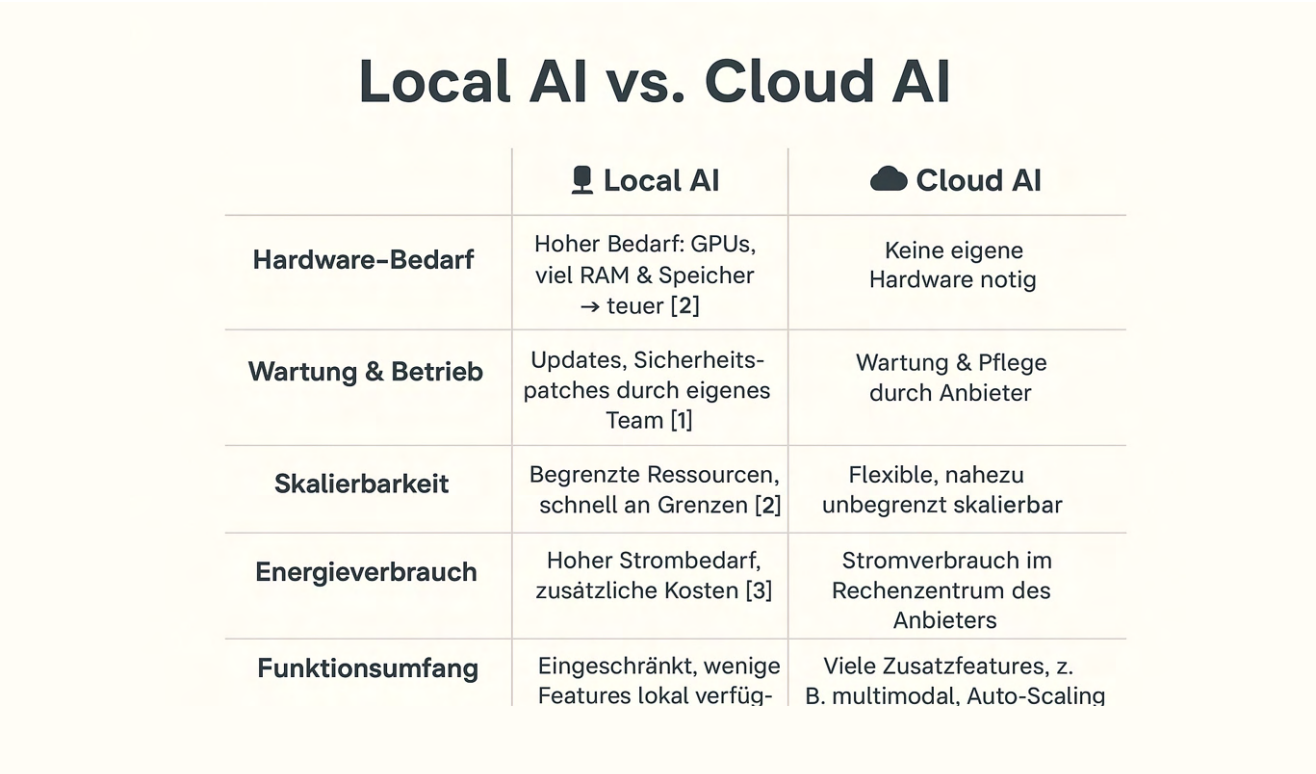


Abbildung 1: Quelle: Bild generiert von Canva.com

Wann lohnt sich Local AI?

Local AI eignet sich besonders für Unternehmen, bei denen Datenschutz und Kontrolle im Vordergrund stehen – z.B. im Gesundheitswesen, in Banken oder in Behörden, oder Unternehmen, die auf Wiederholbarkeit und Stabilität hohen Wert legen.[1] Für Firmen ohne hohe Sicherheitsanforderungen oder ohne große Hardware-Ressourcen ist die Cloud oft praktischer und günstiger.

Fazit

Local AI ist kein Ersatz für Cloud-Angebote, sondern eine Ergänzung. In vielen Szenarien wird ein Hybrid-Ansatz die beste Lösung sein: Vertrauliche Daten werden lokal verarbeitet, während allgemeine Anwendungsfälle in der Cloud bleiben. So profitieren Unternehmen von beiden Welten.[2]

Quellen und weiterführende Informationen:

[1] Heise Online: „Lokale KI-Modelle im Einsatz“ (2024)

[2] Gartner Report: „AI Infrastructure Trends“ (2023)

[3] Open Source Initiative: „On-Premise AI and Privacy“ (2023)

Ahmed Soliman ist Consultant bei SEQIS.

Mit einem Fundament im Maschinenbau und Business Administration fand er seinen Weg in die Welt des Software-Testings und Qualitätsmanagement. Ihn begeistert besonders die Verbindung von technischer Tiefe und strategischer Prozessoptimierung. Seine Schwerpunkte liegen in der Testautomatisierung sowie der Anwendung moderner Tools wie Playwright. Er verfolgt das Ziel, durch kontinuierliche Weiterentwicklung und den Einsatz agiler Methoden die Softwarequalität in funktionsübergreifenden Teams nachhaltig zu steigern.

Tschüss Cloud, Hallo Souveränität: die Technologie hinter Local AI

von Martin Brandhuber

Künstliche Intelligenz / Artificial Intelligence überflutet alle Märkte. Überall findet man AI - im Kinderspielzeug genauso wie im Navigationssystem des neuen Autos, in medizinischen Betrieben, Industriefabriken, Zentren für Forschung und Entwicklung, und natürlich auch in der Weiterbildung. Je mehr AI unser Leben und die dahinter laufenden digitalen Prozesse durchdringt, desto weniger haben wir die Kontrolle darüber. Fragen, die wir uns im Kontext stellen sollten:. Welche AI sieht sich meine Daten an und zu welchem Zweck? Wo speichert diese AI die gesammelten Daten? Was mache ich, wenn "meine eigene" AI in der Cloud meine Daten herumreicht? Jahrelang war die Cloud die einzig wahre Lösung - losgelöst von allen Einschränkungen was Rechenleistung, Datenverfügbarkeit und Energieverbrauch betrifft, war es oft auch nur "cool und hip", alles in die Cloud zu verlagern. Aber spätestens mit dem Boom der AI ändert sich der Fokus: wie kann ich sicherstellen, dass eine AI nicht irgendwo in der Welt herumläuft und meine Daten missbraucht? Wie kann ich die teils horrenden Kosten besser steuern, die eine Cloudlösung mit sich bringt?

Hier betritt Local AI die Bühne.

I. Strategische Positionierung: Warum Local AI für Unternehmen relevant ist

Local AI bezieht sich auf Künstliche Intelligenz-Modelle und -Anwendungen, die direkt auf einem Endgerät oder der lokalen Infrastruktur (wie z. B. einem Laptop, Desktop-PC, Smartphone, IoT-Gerät oder einem Server im eigenen Unternehmen) ausgeführt werden, anstatt sich auf externe Cloud-Server zu verlassen. Es ist das Gegenstück zur Cloud AI, bei der die Datenverarbeitung und die Inferenz (die Anwendung des Modells) auf entfernten Rechenzentren wie denen von OpenAI, Google oder Amazon Web Services stattfinden. Unternehmen, die AI nutzen möchten, stehen nun vor der grundlegenden Entscheidung, wo ihre AIs laufen sollen: in der Cloud oder lokal. Die Daten im eigenen Haus zu haben ist natürlich nicht der einzige Point of Interest: die steigende Bedeutung von lokalen AI-Modellen wird getrieben durch fundamentale Anforderungen an Geschwindigkeit, Kontrolle und Kostenstruktur.

Wann kommt Local AI zum Einsatz?

Local AI ist besonders vorteilhaft für Anwendungsfälle, bei denen Datenschutz, Echtzeit-Reaktion und Unabhängigkeit kritisch sind:

- Datenschutz: In Branchen wie Gesundheitswesen, Finanzen und Rechtswesen, wo vertrauliche Daten das Unternehmensnetzwerk nicht verlassen dürfen (z. B. lokale Analyse von Patientendaten).
- Echtzeit-Anwendungen: Bei autonomen Fahrzeugen, industrieller Automatisierung oder Überwachungssystemen, die sofortige Entscheidungen ohne Verzögerung benötigen.
- Offline-Nutzung: Mobile Anwendungen, die in Gebieten mit schlechter oder keiner Internetverbindung funktionieren müssen (z. B. Übersetzungen, Bilderkennung).
- Kostenkontrolle: Unternehmen mit sehr hohem und konsistenten Nutzungsvolumen, für welche die Pay-per-Use-Gebühren der Cloud zu teuer werden.

Abbildung 1: Quelle: Bild generiert von Gemini (Google AI)

I.a. Definition und Kernvorteile der lokalen AI

Local AI bedeutet, dass die gesamte Datenverarbeitung direkt auf den lokalen Geräten oder Servern stattfindet. Dadurch werden die Kommunikationsverzögerungen und die Abhängigkeit von externen Cloud-Diensten vollständig eliminiert. Ein entscheidender Vorteil ist die extrem niedrige Latenz und damit die Echtzeitfähigkeit. Da keine Daten zur Verarbeitung in ein externes Rechenzentrum und zurück gesendet werden müssen, ist die lokale Ausführung signifikant schneller. Dies ist kritisch für industrielle Anwendungen, bei denen eine Reaktion innerhalb von Millisekunden erfolgen muss, um Betriebssicherheit und Effizienz zu gewährleisten. Die zweite Säule der lokalen AI ist die Datensouveränität und Kontrolle. Local AI stellt sicher, dass

Unternehmen die vollständige Kontrolle über ihre AI-Modelle und alle verarbeiteten Daten behalten. Dies ist essenziell für den Schutz sensibler Informationen und die Einhaltung strenger Datenschutzvorschriften. Im Cloud-Modell mietet man Rechenressourcen, wobei die Datenkontrolle beim Cloud-Anbieter verbleibt. In Bezug auf die Kostenstruktur verschiebt Local AI die Ausgaben, wenn man diese beschafft: statt dynamischer, wiederkehrender Betriebskosten fallen höhere einmalige Anfangsinvestitionen an. Die fortlaufenden Gebühren für Cloud-Speicherung, Datentransfer und Rechenleistung entfallen, da die Verarbeitung intern erfolgt. Viele aktuelle AI-Entwicklungen konzentrieren sich darauf, große Modelle zu optimieren, damit sie auch auf normaler Consumer-Hardware wie modernen Laptops effizient laufen können.

I.b. Skalierbarkeit und Hybride Architekturen

Die größte technische Herausforderung von Local AI ist die Skalierbarkeit. Während Cloud-Dienste dynamisch Kapazitäten mieten können, erfordert Local AI physische Hardware-Anpassungen, was komplexer und langsamer ist. Daher sind hybride Architekturen oft eine gute Lösung. Sie nutzen die Vorteile beider Welten: Local AI sorgt für Datenschutz und Echtzeitverarbeitung, während Cloud AI Flexibilität und dynamische Skalierbarkeit für weniger latenzkritische Aufgaben bietet. Ein Szenario wäre etwa, dass die lokalen Geräte zeitkritische Entscheidungen selbst treffen, während die Cloud zur zentralen Optimierung der Modelle (z. B. durch das Training mit Daten) genutzt wird.

Strategischer Vergleich: Local AI vs. Cloud AI

Kriterium	Local AI / On Premise	Cloud AI / API-basiert
Latenz	Extrem niedrig, ideal für Echtzeit-Anwendungen	Abhängig von Netzwerk; höher
Datenkontrolle & Datenschutz	Vollständige Kontrolle; Daten bleiben lokal	Daten werden an Drittanbieter übertragen; Compliance-Risiko
Kostenstruktur	Hohe initiale Investitionen (CAPEX); niedrige laufende Kosten	Niedrige initiale Investition; hohe, dynamische Betriebskosten (OPEX)
Skalierbarkeit	Erfordert physische Hardware-Anpassungen	Dynamisch und elastisch; Kapazitäten einfach mietbar

II. Die Kunst der Miniaturisierung: Modelloptimierung für lokale Hardware

Moderne AI-Modelle, insbesondere Large Language Models (LLMs), sind oft von gigantischen Ausmaßen. Da lokale Hardware begrenzt ist, muss die Modellgröße demzufolge drastisch reduziert werden. Diese Miniaturisierung ist der Schlüssel zur praktischen Umsetzung von Local AI.

II.a. Pruning: Die Entfernung von Redundanzen

Beim Pruning (Beschneiden) wird das neuronale Netz „ausgemistet". Es basiert auf der Erkenntnis, dass viele Parameter in einem AI-Modell unnötig sind, weil sie nicht oder kaum zur Entscheidungsfindung beitragen. Pruning entfernt diese unnötigen Teile, wodurch das Modell kleiner, schneller und effizienter wird. Man unterscheidet zwischen

- Unstrukturiertes Pruning: Entfernt einzelne, unwichtige Parameter (Gewichte). Das Ergebnis ist zwar sehr kompakt, erfordert aber spezialisierte Hardware, um wirklich schneller zu laufen.
- Strukturiertes Pruning: Entfernt ganze Komponenten (wie komplette Neuronen oder Layer). Dies ist hardwarefreundlicher, da die resultierende Architektur besser mit Standard-GPUs zusammenarbeitet und die Geschwindigkeit direkt verbessert.

II.b. Quantisierung: Die Vereinfachung der Zahlen

Quantisierung reduziert die Präzision der im Modell verwendeten Zahlen. AI-Modelle nutzen standardmäßig hochpräzise 32-Bit-Zahlen (FP32), was viel Speicher und Rechenzeit benötigt. Quantisierung rundet diese auf kleinere, einfachere Repräsentationen (z.B. 8-Bit- oder 4-Bit-Ganzzahlen) ab. Dies senkt den Speicherbedarf und den Rechenaufwand drastisch.

Der Kompromiss liegt zwischen Kompression und Genauigkeit:

- 8-Bit Quantization (INT8): Reduziert den Speicherbedarf bei nur minimalem Genauigkeitsverlust. Dies ist ideal für Server-Setups, bei denen hohe Präzision erforderlich ist
- 4-Bit Quantization (INT4): Ermöglicht eine umfassende Spei-

chereinsparung und beschleunigt die Verarbeitung massiv. Der Nachteil ist ein merklicherer Genauigkeitsabfall - dies wird für Edge-Geräte oder Consumer-GPUs mit sehr begrenztem Speicher genutzt.

Quantisierungsstrategien

- Post-Training Quantization (PTQ): Die Quantisierung wird auf ein bereits fertig trainiertes Modell angewandt. Dies ist die schnellste und einfachste Methode, kann aber einen Genauigkeitsverlust verursachen.
- Quantization-Aware Training (QAT): Die Quantisierungsoperationen werden in den Trainingsprozess integriert. Das Modell lernt, sich auf die reduzierte Präzision einzustellen. QAT liefert bessere Ergebnisse, erfordert aber signifikant mehr Rechenressourcen und Zugang zu repräsentativen Trainingsdaten.

Mechanismen der Modellkompression und Trade-offs[1]

Technik	Funktionsweise	Speichereinsparung (LLM-Fokus)	Genauigkeit (Trade-off)
4-Bit Quantisierung	Reduziert Präzision auf 4 Bits (INT4)	Bis zu 75% Reduktion	Genauigkeitsverlust
8-Bit Quantisierung	Reduziert Präzision auf 8 Bits (INT8)	Etwa 50% Reduktion	Genauigkeitsverlust
Strukturiertes Pruning	Entfernt ganze Komponenten (Neuronen/Heads)	Variabel (bis zu 50% in LLMs)	Beibehaltung (bei 50% Reduktion)
Knowledge Distillation	Kleines Modell lernt von großem Modell	Reduziert Modellgröße signifikant	Kann ähnliche Genauigkeit wie der Lehrer erreichen

II.c. Knowledge Distillation (KD): Der Lehrer-Schüler-Ansatz

Knowledge Distillation ist eine Strategie, bei der das Wissen von einem großen, leistungsstarken „Lehrer"-Modell auf ein kleineres, effizienteres „Schüler"-Modell übertragen wird. Der Schüler lernt, die Ausgaben des Lehrers nachzuahmen. Dadurch kann er eine ähnliche oder höhere Genauigkeit erzielen, ohne so viel Rechenkapazität zu beanspruchen. Dies ist ein kosteneffektives Verfahren, um leistungsstarke, aber kleinere Modelle aus einem initialen, größeren Geschwistermodell abzuleiten.

III. Die Laufzeitumgebung: Inference Engines und Software-Stack

Nach der Optimierung muss das AI-Modell effizient auf der Zielhardware ausgeführt werden. Hierfür sind spezialisierte Software-Infrastrukturen – die Inferenz-Engines – notwendig. Sie sorgen für niedrige Latenzzeiten und einen hohen Durchsatz (gemessen in Tokens pro Sekunde).

III.a. Modell-Compiler und Interoperabilität

Modelle aus Trainings-Frameworks sind selten direkt für die Inferenz auf lokalen Geräten optimiert. Sie müssen kompiliert werden, um die maximale Effizienz der lokalen Hardware zu nutzen. Wichtige Formate und Laufzeitumgebungen (Compiler) hierfür sind etwa das ONNX-Format (Open Neural Network Exchange), das von Engines wie NVIDIAs TensorRT LLM genutzt wird, oder TFLite und Edge AI Suites. Die Inferenz-Engine ist die kritische Komponente, da sie die hardwarespezifischen Optimierungen vornimmt und somit die Geschwindigkeit der lokalen AI verbessert.

III.b. Verteilte und skalierbare Lokale Inferenz

Um die Skalierbarkeitsprobleme lokaler Hardware zu umgehen, nutzen fortschrittliche Engines Peer-to-Peer (P2P) Netzwerke zur sicheren und privaten Verteilung von Inferenzanfragen über mehrere lokale Server hinweg.

Zwei Modi der verteilten Inferenz sind wichtig:

- Federated Mode (Data Parallelism / Datenparallelität): Anfragen werden vom Lastverteiler an einen einzigen Worker Node im Cluster weitergeleitet. Dies dient der einfachen Lastverteilung über mehrere gleichartige Server.
- Worker Mode (Model Sharding / Modellparallelität): Das AI-Modell wird in Teile zerlegt und auf verschiedene Worker Nodes aufgeteilt. Alle Worker bearbeiten die Anfrage gemeinsam. Dies ist die

Lösung, um Modelle, die zu groß für eine einzelne GPU sind, dennoch lokal betreiben zu können.

IV. Der Hardware-Engpass: Spezielle Anforderungen an lokale Infrastruktur

Bei Local AI diktiert die physikalische Größe der Modellgewichte die Hardware-Wahl. Der entscheidende Engpass ist nicht die reine Rechenleistung, sondern der Speicher und dessen Geschwindigkeit (Bandbreite).

Der Dreiklang: CPU, GPU und NPU

- GPU (Graphics Processing Unit): Der Standardbeschleuniger für das Training großer Modelle und die Inferenz mit hohem Durchsatz.
- NPU (Neural Processing Unit): Speziell entwickelte Chips, optimiert für AI-Inferenz mit maximaler Energieeffizienz und niedrigstem Stromverbrauch.
- CPU (Central Processing Unit): Dient als Fallback. Sie kann sehr große Modelle ausführen, die nicht in den VRAM passen, indem sie auf den langsameren System-RAM zugreift. Dies ist ein langsamer, aber notwendiger Notfall-Betriebsmodus.

V.a Der Deployment-Workflow für lokale AI

Die Implementierung von Local AI folgt einem strukturierten Prozess und eröffnet spezifische Anwendungsfelder:

- Datenerfassung und -vorbereitung.
- Modelltraining und Fine-Tuning (Anpassung des Modells an spezifische Unternehmensdaten).
- Optimierung: Reduzierung von Größe und Ressourcenverbrauch durch Quantisierung und Pruning.
- Kompilierung und Validierung: Das optimierte Modell wird für die Zielhardware kompiliert.
- Bereitstellung: Ausrollen des kompilierten Modells auf die lokale Infrastruktur oder Endgeräte.

V.b. Geschäftliche Anwendungsfälle im Industrial IoT (IIoT)

Local AI ist ein wesentlicher Treiber für die digitale Transformation in der Industrie:

- Echtzeit-Entscheidungsfindung: Die lokale Verarbeitung von Daten in Produktionsumgebungen reduziert Reaktionszeiten und ist entscheidend für die Sicherheit und Automatisierung.
- Dateneffizienz: Durch lokale Filterung und Komprimierung der Rohdaten reduziert Edge Computing das zu übertragende Datenvolumen signifikant. Nur die vorverarbeiteten, relevanten Informationen werden an höhere Systeme gesendet, was Netzwerklasten und Kommunikationskosten senkt.

VI. Ausblick und zukünftige technologische Treiber

Die technologische Entwicklung in Local AI zielt darauf ab, die Leistung lokal betriebener LLMs weiter zu steigern, insbesondere durch die Überwindung des Speicherengpasses.

VI.a. Hardware-Innovationen zur Überwindung des Speicherengpasses

- Unified Memory Architekturen: Zukünftige Systeme werden einen hochintegrierten Speicherpool verwenden. Obwohl die Speicherbandbreite weiterhin der Hauptengpass bleibt, ermöglicht diese Integration die Ausführung von Modellen, die auf herkömmlichen Desktop-Systemen nicht funktionieren würden.
- Architekturwandel: Im Edge Computing verschieben sich die Architekturen von General Purpose GPUs hin zu anwendungsspezifischen integrierten Schaltungen (ASICs), die speziell für AI-Workloads optimiert sind. Ein ASIC kann nur die eine Aufgabe, für die er gebaut wurde. Er kann nichts anderes. Aber diese eine Aufgabe erledigt er unschlagbar schnell und mit extrem niedrigem Energieverbrauch (siehe auch NPUs).

VI.b. Algorithmen für Sparse Computation

Die Rechenkosten steigen stark mit der Größe des Kontextes (Textlänge). "Sparse Attention" ist ein Ansatz, bei dem das Modell nur auf die wichtigsten Teile des Kontexts schaut. Neue Forschung, wie Native Sparse Attention (NSA), integriert diese algorithmischen Verbesserungen mit hardware-optimierten Implementierungen. Dies ermöglicht Geschwindigkeitssteigerungen und die Verarbeitung von sehr langen Texten in lokal betriebenen LLMs.

Weitere spannende Blogartikel finden Sie hier:

www.seqis.com/de/blog-index

Abbildung 2: Quelle: Bild generiert von Gemini (Google AI)

VII. Zusammenfassung und strategische Implikationen

Local AI ist die beste Lösung für Unternehmen, die absolute Datenkontrolle, regulatorische Compliance und minimale Latenz benötigen. Technologisch basiert die Machbarkeit auf drei Säulen:

1. Miniaturisierung: Quantisierung (Zahlen vereinfachen) und Pruning (unnötige Teile entfernen) sind notwendig, um große Modelle auf bezahlbarer Hardware lauffähig zu machen.
2. Inferenz-Engines: Spezialisierte Software (z.B. TensorRT) maximiert die Leistung der Hardware
3. Hardware-Priorität: Bei der Beschaffung sind die Speicherkapazität und die Speicherbandbreite wichtiger als die reine Rechenleistung, da diese die maximale Modellgröße und die effektive Geschwindigkeit bestimmen.

Dieser Artikel stellt nur einen kleinen Einblick in die Technologie hinter Local AI dar. Auf den ersten Blick mag diese technologische Tiefe, von der komplexen Modelloptimierung bis hin zur Verwaltung verteilter Infrastrukturen, unglaublich komplex wirken und hohe Anforderungen an die interne IT stellen. Dennoch ist Local AI nicht nur eine Nischenlösung, sondern wird für viele Unternehmen, die AI strategisch und datenschutzkonform nutzen möchten, in Zukunft ein zu überlegender Faktor sein. Wer die AI-gestützte digitale Transformation anführen und gleichzeitig die Kontrolle über seine kritischsten Assets behalten will, muss sich mit der Komplexität der Local AI auseinandersetzen, um die Wettbewerbsfähigkeit des Unternehmens langfristig zu sichern.

Quellen und weiterführende Informationen:

[1] https://www.newline.co/@zaoyang/4-bit-vs-8-bit-quantization-key-differences--842272c7

https://www.mind-verse.de/post/ki-fuer-pruning-neuronale-netzwerk-optimierung

https://medium.com/@shmilysyg/model-compression-pruning-quantization-distillation-and-binarization-7710ac954567

https://medium.com/@yugank.aman/knowledge-distillation-for-llms-techniques-and-applications-e23a17093adf

Martin Brandhuber ist Senior Consultant und Teamlead bei SEQIS.

„Kundenzufriedenheit und Professionalität in der Planung und Durchführung des Testprozesses hat für mich höchste Priorität. Testen ist für mich ein nicht wegzudenkender Bestandteil der Softwareentwicklung. Da die Anforderungen an den Vernetzungsgrad von Systemen laufend zunehmen, ist die frühe Einbindung von Testprozessen in der Softwareentwicklung unabdingbar. Nur so kann die Entwicklung kostengünstiger, qualitativ hochwertiger Software garantiert werden."

razzfazz.ai - local matters

von Alexander Vukovic

1. Einführung: Die Renaissance der lokalen Intelligenz

Wir leben in einer Ära der digitalen Paradoxien. Auf der einen Seite erleben wir durch generative künstliche Intelligenz (GenAI) einen technologischen Sprung, der in seiner Tragweite mit der Erfindung des Internets vergleichbar ist. Modelle werden größer, schneller und fähiger. Auf der anderen Seite zieht sich das regulatorische Netz enger. Der **EU AI Act** ist nicht länger ein fernes bürokratisches Gespenst, sondern gebundene Realität, die Unternehmen dazu zwingt, ihre KI-Strategien grundlegend zu überdenken. In der aktuellen Ausgabe der SEQIS Quality News 2025-2, die unter dem Titel „AI Act & Local AI – Wege in eine verantwortungsvolle Zukunft" steht, wird dieser Spannungsbogen deutlich skizziert. Es geht nicht mehr nur darum, was technisch machbar ist, sondern was rechtlich zulässig, ethisch vertretbar und ökonomisch sinnvoll ist.

Als jemand, der die Software-Qualitätssicherung seit den frühen Tagen der agilen Bewegung begleitet hat – von den ersten Gehversuchen mit Testautomatisierung bis hin zu komplexen Continuous-Integration-Pipelines –, sehe ich Parallelen. Damals ging es darum, Silos zwischen Entwicklung und Betrieb aufzubrechen (DevOps). Heute müssen wir das Silo zwischen „mächtiger KI in der Cloud" und „sicheren Daten on-premise" aufbrechen. Die Antwort darauf lautet nicht, Daten blind in die Cloud zu schieben und auf Verschlüsselung zu hoffen. Die Antwort lautet: Die Intelligenz muss zu den Daten kommen.

Mit **razzfazz.ai** haben wir bei SEQIS eine Plattform geschaffen, die genau diese Philosophie verkörpert: „Enabling local AI". Es ist ein Manifest für digitale Souveränität. Warum ist das gerade jetzt so kritisch? Weil die Technologie endlich so weit ist. Wir müssen nicht mehr wählen zwischen „dumm und lokal" oder „schlau und Cloud". Durch die Symbiose aus spezialisierter Hardware (Unified Memory Architecture) und hocheffizienten Open-Source-Modellen wie **Magistral** und **Devstral** von Mistral AI können wir Enterprise-Grade-Intelligence lokal betreiben. razzfazz.ai ist also keine eigene "KI" sondern eine lokale Hardware + Softwareplattform um beliebige Open Source KI-Modelle laufen lassen zu können.

Dieser Artikel ist ein technischer Deep Dive, eine Bestandsaufnahme der Möglichkeiten. Wir werden analysieren, warum lokale Hardware plötzlich wieder „sexy" ist, wie Reasoning-Modelle die Testanalyse revolutionieren und wir werden zehn konkrete, implementierbare Use-Cases durchspielen, die zeigen, wie Entwicklung und Testing im Jahr 2026 funktionieren sollten.

2. Der technologische Unterbau: Hardware und Architektur

Um zu verstehen, warum lokale KI heute eine viable Alternative zu Cloudangeboten wie OpenAI oder Anthropic darstellt, müssen wir einen Blick unter die Haube werfen. Die traditionelle PC-Architektur war für KI-Workloads denkbar ungeeignet. Der Flaschenhals war immer der Datentransfer zwischen der CPU (Central Processing Unit) und der GPU (Graphics Processing Unit) über den PCIe-Bus.

2.1 Unified Memory: Der Gamechanger

Die razzfazz.ai Box setzt auf eine Architektur, die diesen Flaschenhals eliminiert: **Unified Memory**. In dieser Konfiguration teilen sich CPU und GPU denselben physischen Arbeitsspeicherpool. Dies hat massive Implikationen für die Inferenz von Large Language Models (LLMs):

1. **Kein Kopier-Overhead:** Bei herkömmlichen Systemen müssen die Gewichte des neuronalen Netzes (die Model Weights) in den VRAM der Grafikkarte geladen werden. Ist das Modell zu groß für den VRAM (z.B. 24 GB bei einer Consumer RTX 4090), muss „Offloading" betrieben werden – Teile des Modells werden auf den langsamen System-RAM ausgelagert, was die Performance dramatisch einbrechen lässt. Mit Unified Memory greift die GPU direkt auf den gesamten Speicher zu.

2. **Kapazität für große Modelle:** Die razzfazz.ai Box bietet bis zu **128 GB Unified Memory**, wovon netto **96 GB** exklusiv für KI-Modelle nutzbar sind. Das ermöglicht den Betrieb von Modellen, die weit über die Kapazitäten klassischer Desktop-GPUs hinausgehen. Wir können hier quantisierte Versionen von Modellen mit 70 Milliarden Parametern oder mehr laden, ohne Leistungseinbußen hinnehmen zu müssen.

3. **Energieeffizienz:** Ein oft unterschätzter Faktor, besonders im Hinblick auf ESG-Ziele (Environmental, Social, and Governance). Eine High-End-Server-GPU kann unter Last 400 bis 700 Watt verbrauchen. Ein Cluster davon benötigt eigene Klimatisierung. Die razzfazz.ai Hardware operiert in einem Bereich von **30 bis 300 Watt.** Dies ist nicht nur gut für die Stromrechnung, sondern ermög-

licht den Einsatz in normalen Büroumgebungen ohne dedizierten Serverraum.

In der folgenden Tabelle vergleichen wir die Architekturansätze, um die Positionierung der lokalen Lösung zu verdeutlichen:

Merkmal	Klassische Workstation + GPU	Cloud API (z.B. GPT-5)	razzfazz.ai Box (Local Unified Memory)
Speicherarchitektur	Getrennt (RAM + VRAM)	Opak / Verteilt	Unified (Shared RAM)
Max. Modellgröße	Begrenzt durch VRAM (z.B. 24 GB)	Sehr groß (Black Box)	Bis zu 96 GB (nutzbar)
Latenz	Niedrig (lokal)	Hoch (Netzwerkabhängig)	Sehr niedrig (lokal)
Datenschutz	Hoch (lokal)	Risiko (Drittanbieter)	Maximum (lokal + air-gapped möglich)
Energieverbrauch	Hoch (800W+)	Unbekannt (extern)	Effizient (30-300W)
Investitionsart	CAPEX (Hardware)	OPEX (Pay-per-Token)	CAPEX (einmalig)

2.2 Der Software-Stack: Orchestrierung der Intelligenz

Hardware ist nur so gut wie die Software, die sie steuert. Die razzfazz.ai Distribution basiert auf einem robusten Linux-Unterbau (Ubuntu 24.04 LTS) und integriert eine Suite von Open-Source-Tools, die nahtlos zusammenarbeiten.

- **llama.cpp:** Dies ist der Motor. Das Projekt hat die Demokratisierung von LLMs massiv vorangetrieben. Es ermöglicht die hocheffiziente Ausführung von GGUF-quantisierten Modellen auf Apple Silicon und anderen Unified-Memory-Architekturen. Durch Quantisierung (z.B. auf 4-bit oder 8-bit Integers) wird der Speicherbedarf massiv reduziert, bei vernachlässigbarem Qualitätsverlust.
- **Workflow Automationen:** Das Nervensystem. Während viele KI-Interaktionen noch über Chat-Fenster laufen, liegt die wahre Macht in der Automatisierung. Mittels unterschiedlicher open source Workflow Automationstools werden KI-Modelle in komplexe Abläufe eingebunden. Man kann es sich als „Klebstoff" vorstellen, der die KI mit Jira, Git, Datenbanken oder E-Mail-Servern verbindet – und das alles lokal.
- **Vektordatenbank (PostgreSQL + pgVector):** Das Langzeitgedächtnis. Um RAG (Retrieval Augmented Generation) lokal umzusetzen, benötigen wir einen Speicher für semantische Embeddings. PostgreSQL läuft effizient auf der Box und ermöglicht es der KI, „Wissen" firmeninterner Dokumente abzurufen, ohne dass das Modell neu trainiert werden muss.
- **Open WebUI:** Die Schnittstelle zum Menschen. Eine benutzerfreundliche Oberfläche, die sich wie ChatGPT anfühlt, aber vollständig lokal läuft und Multi-User-Support sowie Modell-Management bietet.

3. Die Modelle: Magistral und Devstral

Die Hardware stellt die Arena, aber die Modelle sind die Athleten. Wir konzentrieren uns in diesem Artikel auf zwei spezifische europäische Modelle von Mistral AI, die für Entwicklung und Testing besonders relevant sind: **Magistral und Devstral**. Beide repräsentieren den neuesten Stand

der Open-Weights-Forschung und sind für den lokalen Betrieb optimiert.

3.1 Magistral: Der Analyst mit Tiefgang (Reasoning Model)
„Magistral" markiert einen Paradigmenwechsel bei Mistral AI. Es ist das erste dedizierte **Reasoning-Modell** des Unternehmens. Anders als klassische LLMs, die primär darauf trainiert sind, statistisch wahrscheinliche Wortfolgen zu generieren, wurde Magistral darauf optimiert, Probleme durch eine „Kette von Gedanken" (Chain-of-Thought, CoT) zu lösen.

Was bedeutet „Reasoning" technisch? Wenn Sie Magistral eine komplexe Frage stellen, antwortet es nicht sofort. Es generiert zunächst einen internen Monolog (oft sichtbar in <think> Tags), in dem es das Problem zerlegt, Hypothesen aufstellt, diese prüft und erst dann das Endergebnis formuliert. Dieser Prozess ähnelt dem „System 2" Denken nach Daniel Kahneman – langsam, logisch, berechnend.

Leistungsdaten und Architektur:

- **Größe:** Magistral Small ist ein 24B (24B = 24 billions, 24 Milliarden) Parameter Modell. Das ist der „Sweet Spot" für lokale Inferenz auf unserer Hardware. Es ist klein genug, um schnell zu sein, aber groß genug für komplexe Logik.
- **Kontext:** Es unterstützt ein Kontextfenster von 128k Token. Das bedeutet, es kann hunderte Seiten Dokumentation oder umfangreiche Code-Dateien im Arbeitsspeicher halten und in seine Überlegungen einbeziehen.
- **Benchmarks:** Auf dem AIME2024 Benchmark (Mathematik und Logik) erreicht die Medium-Variante Scores von über 73%, was eine massive Steigerung gegenüber nicht-reasoning Modellen darstellt.
- **Anwendungsgebiet:** Software-Architektur, Requirements Engineering, logische Prüfung von Testfällen, komplexe Datenanalyse.

Ein interessantes Feature ist die Geschwindigkeit bei einfachen Aufgaben: Durch „Flash Answers" kann das Modell bei weniger komplexen Anfragen extrem schnell antworten, skaliert aber bei Bedarf seine Rechenzeit für tiefere Analysen hoch.

3.2 Devstral: Der Agentische Entwickler (Coding Model)
Während Magistral der Denker ist, ist **Devstral** der Macher. Es wurde spezifisch für **Software Engineering** Aufgaben trainiert und verfeinert. Es ist nicht einfach nur eine Autovervollständigung; es ist für „Agentic Workflows" konzipiert.

Der Unterschied zu klassischen Code-Modellen:
Frühere Modelle wie CodeLlama waren gut darin, eine Funktion zu vervollständigen. Devstral hingegen versteht den Kontext eines gesamten Projekts. Es wurde in Zusammenarbeit mit All Hands AI entwickelt und darauf trainiert, wie ein Entwickler zu agieren: Es kann Dateien lesen, Änderungen planen, Code editieren und Shell-Befehle ausführen (in einer gesicherten Sandbox).

Spezifikationen:

- **Spezialisierung:** Es basiert auf Mistral Small 3.1, wobei der Vision-Encoder entfernt wurde, um sich rein auf Text und Code zu konzentrieren.
- **Performance:** Im **SWE-Bench Verified**, dem Goldstandard für automatisches Software-Engineering, erreicht Devstral einen Score von **46.8%**. Zum Vergleich: GPT-4o mini liegt bei ca. 23.6%. Das bedeutet, Devstral löst fast die Hälfte aller realen GitHub-Issues autonom korrekt – ein Wert, der für ein Modell dieser Größe (24B) sensationell ist.
- **Tokenizer:** Es nutzt den Tekken-Tokenizer mit einem Vokabular von 131k, was besonders effizient für Source Code ist (weniger Token pro Codezeile = mehr Kontext im Speicher).

Zusammen bilden Magistral und Devstral ein unschlagbares Team: Der eine plant und prüft (Magistral), der andere führt aus und implementiert (Devstral).

4. Zehn Use-Cases für Entwicklung und Testing

Theorie ist gut, Praxis ist besser. Wie setzen wir diese PS auf die Straße? Im Folgenden skizziere ich zehn detaillierte Use-Cases, die wir bei SEQIS nutzen und die mit der razzfazz.ai Box umsetzbar sind. Jeder Use-Case nutzt die spezifischen Stärken von Magistral (Reasoning) oder Devstral (Coding) und die Orchestrierung durch Workflow Automation.

Use-Case 1: Der Requirements-Architekt (Modell: Magistral)

Herausforderung:
Anforderungen in der Softwareentwicklung sind oft unpräzise. Lastenhefte bestehen aus Prosatext, E-Mails und Meeting-Notizen. Die Überführung in formale Modelle (wie UML) ist ein manueller, fehleranfälliger Prozess. Missverständnisse hier führen zu den teuersten Fehlern später im Projekt.

Die Lösung mit lokaler AI:
Wir nutzen Magistral, um unstrukturierte Anforderungen in formale Diagramme zu übersetzen.

1. **Input:** Der Nutzer lädt ein PDF-Lastenheft oder kopiert Text in das Open WebUI.
2. **Reasoning-Prozess:** Ein automatisierter Workflow übergibt den Text an Magistral mit dem System-Prompt: „Analysiere diesen Text auf Akteure, Use Cases und Ablauflogik. Identifiziere logische

Lücken (z.B. fehlende Fehlerbehandlung)."

3. **Generierung:** Magistral generiert validen Code für PlantUML oder Mermaid (z.B. für ein Aktivitätsdiagramm). Dank seiner Reasoning-Fähigkeit kann es implizite Annahmen explizit machen (z.B. „Wenn der User nicht eingeloggt ist, muss er zum Login geleitet werden", auch wenn das im Text fehlt).
4. **Visualisierung:** Mittels Workflow Automation wird der PlantUML-Code direkt in eine Grafik gerendert und diese zusammen mit einer Liste der identifizierten Lücken zurückgeliefert.

Mehrwert:
Standardisierung der Dokumentation und Qualitätssicherung der Anforderungen bevor eine Zeile Code geschrieben wird („Shift Left"). Da Anforderungen oft sensible Geschäftsstrategien enthalten, ist der lokale Betrieb essenziell.

Use-Case 2: Der Gnadenlose Test-Designer (Modell: Magistral)

Herausforderung:
Testfallerstellung ist oft eine monotone Arbeit. Tester neigen dazu, den „Happy Path" zu testen und Randfälle zu übersehen. Kombinatorische Explosionen (z.B. bei komplexen Formularen) überfordern menschliche Tester oft.

Die Lösung mit lokaler AI:
Der „Testing Buddy" nutzt Magistrals logische Tiefe.

1. **Input:** Eine User Story oder die Ergebnisse aus Use-Case 1.
2. **Analyse:** Magistral wendet formale Testmethoden an, wie Äquivalenzklassenbildung und Grenzwertanalyse. Es „denkt" sich Szenarien aus, die das System brechen könnten (Destructive Testing).
3. **Workflow:**
   - Magistral identifiziert alle Eingabefelder.
   - Es generiert Testdaten für Grenzwerte (z.B. Alter 17, 18, 120, -1).
   - Es erstellt eine Testmatrix.
4. Über die Workflow Automation werden diese Testfälle direkt in das Format von **Jira Xray oder Tricentis Tosca** konvertiert und via API in das Testmanagement-Tool importiert.

Mehrwert:
Massive Zeitersparnis und eine höhere Testabdeckung durch systematische, KI-gestützte Analyse. Die Daten verlassen nie das lokale Netzwerk.

Use-Case 3: Legacy Code Refactoring & Modernisierung (Modell: Devstral)

Herausforderung:
Viele Unternehmen sitzen auf „Legacy Code" – funktionierende, aber veraltete Software (z.B. Java 6, alter PHP-Code), die niemand mehr anfassen will. Das Wissen darüber ist oft mit Mitarbeitern in Pension gegangen.

Die Lösung mit lokaler AI:
Devstral nutzt sein riesiges Kontextfenster (128k), um ganze Dateien oder Module zu verstehen.

1. **Kontext:** Wir laden den Quellcode einer alten Komponente in den Kontext von Devstral.
2. **Agentic Task:** Der Prompt lautet: „Analysiere diesen Code. Erkläre die Business-Logik. Schlage ein Refactoring auf moderne Standards (z.B. Java 21 Records, Streams) vor. Behalte die Logik bei."
3. **Iterativer Prozess:** Devstral schreibt den Code um. Da es agentisch agiert, kann es auch Unit-Tests für den alten Code schreiben, um sicherzustellen, dass der neue Code dasselbe tut (Regressionstesting).
4. **Output:** Ein Diff-File oder ein direkter Commit-Vorschlag im lokalen GitLab.

Mehrwert:
Risikominimierung bei der Modernisierung. Da Legacy-Code oft tiefes Firmen-Know-how enthält, darf dieser keinesfalls in öffentliche Cloud-LLMs geladen werden.

Use-Case 4: Synthetische Testdaten-Generierung (DSGVO-konform) (Modell: Magistral)

Herausforderung:
Für Tests werden realistische Daten benötigt. Produktionsdaten zu anonymisieren ist aufwendig und birgt Re-Identifikationsrisiken. Einfache Zufallsgeneratoren erzeugen oft fachlichen Unsinn (z.B. PLZ passt nicht zum Ort).

Die Lösung mit razzfazz.ai:
Magistral generiert semantisch korrekte, synthetische Daten.

1. **Modellierung:** Wir definieren das Datenschema (z.B. JSON-Struktur für einen „Versicherungsnehmer").
2. **Logik-Injektion:** Wir geben Magistral Regeln: „Erzeuge 50 Datensätze für Kunden aus Österreich. Wenn Bundesland = Wien, dann muss die PLZ mit 1 beginnen. Das Geburtsdatum muss zum Status ‚Pensionist' passen."
3. **Reasoning:** Magistral nutzt seine Logik-Fähigkeiten, um diese Abhängigkeiten einzuhalten (etwas, woran einfachere Modelle oft scheitern).
4. **Persistenz:** Die Workflow Automation schreibt die generierten Daten direkt in die lokale PostgreSQL-Datenbank der Box, von wo aus sie in Testumgebungen injiziert werden können.

Mehrwert:
100% DSGVO-Compliance, da keine echten Daten berührt werden, bei gleichzeitig hoher fachlicher Qualität der Testdaten.

Use-Case 5: Der Autonome Bug-Hunter (Modell: Devstral)

Herausforderung:
Bugs zu finden, die nur unter bestimmten Bedingungen auftreten (z.B. Race Conditions), ist extrem schwierig. Statische Codeanalyse findet Syntaxfehler, aber keine logischen Probleme im Ablauf.

Die Lösung mit razzfazz.ai:
Ein lokaler Agent, der Git-Commits überwacht.

1. **Trigger:** Ein Entwickler pusht Code. Die Workflow Automation detektiert den Change.
2. **Analyse:** Devstral liest die Änderungen (Diff) und die betroffenen Dateien im Volltext.
3. **Reasoning:** Devstral sucht nach logischen Fehlern: „In Zeile 50 wird auf user zugegriffen, aber user könnte null sein, wenn die Datenbankabfrage in Zeile 40 fehlschlägt. Es fehlt ein Null-Check.".
4. **Report:** Der Agent postet diesen Hinweis als Kommentar direkt in den Merge Request im lokalen Git-Server.

Mehrwert:
Ein „Pair Programmer", der nie schläft und jeden Commit reviewt. Die Fehlererkennung passiert bevor der Code in den Master-Branch gemergt wird.

Use-Case 6: Self-Healing Unit Tests (Modell: Devstral)

Herausforderung:
Code ändert sich, Tests brechen. Die Wartung von Unit Tests ist oft teurer als die Entwicklung der Features selbst.

Die Lösung mit razzfazz.ai:
Devstrals Fähigkeiten im Bereich „Agentic Coding".

1. **Szenario:** Ein Build schlägt fehl, weil ein Unit Test rot ist.
2. **Diagnose:** Die Workflow Automation fängt den Fehler aus der CI/CD-Pipeline (z.B. Jenkins) ab und sendet den Stack Trace sowie den geänderten Code an Devstral.
3. **Fix:** Devstral analysiert, warum der Test fehlschlägt. War es eine beabsichtigte Änderung der Logik? Dann muss der Test angepasst werden. War es ein Bug? Dann muss der Code gefixt werden.
4. **Aktion:** Devstral schlägt den korrigierten Code vor. In einer fortgeschrittenen Ausbaustufe kann der Agent den Fix lokal ausführen, die Tests erneut laufen lassen und bei Erfolg den Fix committen.

Mehrwert:
Drastische Reduktion der Wartungsaufwände für Testsuiten und stabilere Builds.

Use-Case 7: Intelligente Log-Analyse und Root Cause Analysis (Modell: Magistral)

Herausforderung:
Wenn in der Produktion oder im Testsystem etwas schiefgeht, müssen Ops-Teams oft Gigabytes an Logfiles durchwühlen. Die Korrelation von Fehlern über verschiedene Microservices hinweg ist für Menschen schwer.

Die Lösung mit razzfazz.ai:
Magistral als forensischer Analyst.

1. **Ingest:** Relevante Log-Ausschnitte (z.B. die 5 Minuten rund um einen Crash) werden an die razzfazz.ai Box übermittelt.
2. **Analyse:** Magistral korreliert Zeitstempel und Fehlermeldungen. Es erkennt Muster: „Der Timeout im Payment-Service (14:00:01) verursachte die Null-PointerException im Order-Service (14:00:02).".
3. **Erklärung:** Das Modell generiert eine Zusammenfassung in natürlicher Sprache für das Ops-Team: „Ursache ist wahrscheinlich eine Überlastung der Datenbank X. Empfohlene Maßnahme: Connection Pool prüfen."

Mehrwert:
Signifikante Reduktion der MTTR (Mean Time To Repair). Das Wissen aus den Logs verlässt dabei nie das gesicherte Netzwerk.

Use-Case 8: Automatisierte Dokumentationspflege (Modell: Devstral)

Herausforderung:
Code-Dokumentation ist fast immer veraltet. Niemand aktualisiert gerne das Wiki oder die Swagger-Definitionen nach einer Code-Änderung.

Die Lösung mit razzfazz.ai:
Ein „Documentation Buddy", der den Codebestand überwacht.

1. **Scan:** Devstral iteriert regelmäßig über das Repository.
2. **Vergleich:** Es vergleicht den Code mit der existierenden Dokumentation (z.B. Markdown-Files oder Docstrings).
3. **Update:** Bei Abweichungen generiert Devstral aktualisierte Beschreibungen der Klassen, Methoden und API-Endpunkte. Es kann sogar Diagramme (siehe Use-Case 1) aktualisieren.
4. **Commit:** Die aktualisierte Doku wird als Pull Request eingereicht.

Mehrwert:
Die Dokumentation ist immer „live" und synchron mit dem Code. Das Onboarding neuer Entwickler wird massiv beschleunigt.

Use-Case 9: Der Agile Ticket-Assistent (Agile Buddy)

Herausforderung:
In Dailies wird viel besprochen, aber wenig dokumentiert. „Machst du noch schnell das Ticket für den Bug im Login?" – und dann wird es vergessen oder nur rudimentär („Login geht nicht") angelegt.

Die Lösung mit razzfazz.ai:
Der Agile Buddy, integriert in den Kommunikationsfluss.

1. **Input:** Ein Entwickler spricht eine kurze Notiz in die mobile App des Open WebUI oder schreibt einen schnellen Satz in den Chat: „Hey Jira, Bug im Checkout-Prozess. Wenn man als Gast bestellt und PayPal wählt, kommt ein 500er Fehler."
2. **Processing:** Die Box nutzt Speech-to-Text (falls Audio) und dann Magistral zur Strukturierung.
3. **Enrichment:** Magistral fragt fehlende Infos ab oder ergänzt Kontext (z.B. Browser-Version, Priority). Es formuliert eine professionelle Fehlerbeschreibung mit „Steps to Reproduce".
4. **Action:** Über die Workflow Automation wird das Ticket via Jira-API erstellt und dem richtigen Team zugewiesen.

Mehrwert:
Senkung der Hürde für saubere Dokumentation. Bessere Ticketqualität führt zu schnellerer Bearbeitung.

Use-Case 10: Lokaler RAG-Knowledge-Bot für Tester (Magistral + PostgreSQL + pgVector)

Herausforderung:
Tester müssen oft wissen: „Wie war das nochmal mit der Storno-Logik bei Tarif X?" Die Antwort steht irgendwo in 500 PDFs im SharePoint.

Die Lösung mit razzfazz.ai:
Ein lokales RAG-System (Retrieval Augmented Generation).

1. **Indizierung:** Alle Fachkonzepte, Wikis und alten Testpläne werden von der Box eingelesen, in Vektoren umgewandelt und in der lokalen PostgreSQL-Datenbank gespeichert.
2. **Query:** Der Tester fragt im Chat: „Wie verhält sich das System bei Storno nach 14 Tagen?"
3. **Retrieval & Generation:** Die Workflow Automation sucht die relevanten Textstellen in PostgreSQL. Magistral erhält diese als Kontext und formuliert eine präzise Antwort inklusive Quellenangabe („Laut Fachkonzept V2.3, Seite 45...").

Mehrwert:
Das gesamte Firmenwissen ist sofort und im Dialog verfügbar. Da kein Dokument das Haus verlässt (anders als bei ChatGPT Uploads), bleibt das IP (Intellectual Property) geschützt.

5. Implementierung: Die Workflow Automation Architektur

Wie setzen wir diese Use-Cases technisch zusammen? Das Herzstück der Orchestrierung auf der razzfazz.ai Box ist die Workflow Automation mit unterschiedlichen Werkzeugen. Im Gegensatz zu Cloud-Diensten wie Zapier oder n8n läuft diese hier lokal im Docker-Container, direkt neben dem KI-Modell.

Ein typischer Workflow (z.B. für den Bug-Hunter, Use-Case 5) folgt einem klaren Muster, das wir als „Agentic Pattern" bezeichnen können:

1. **Trigger Node:** Horcht auf Ereignisse (Webhook von GitLab, Dateiänderung, Zeitplan).
2. **Data Fetching:** Holt die notwendigen Daten (Code, Logs, Text).
3. **AI Agent Node / Chain:** Hier passiert die Magie. Wir nutzen den „Basic LLM Chain" Node oder spezialisierte Agenten-Knoten. Dieser Knoten kommuniziert über http://localhost:9090 mit dem **llama.cpp** Server.
   - Wichtig: Wir konfigurieren hier Parameter wie temperature (niedrig für Code/Logik, höher für kreative Texte) und den System-Prompt.
4. **Tool Use:** Der Agent kann entscheiden, Tools aufzurufen. In der Workflow Automation können wir dem Agenten „Tools" bereitstellen, z.B. „Calculator", „Database Query" oder „Git Commit". Devstral ist besonders gut darin, zu erkennen, wann es welches Tool nutzen muss.
5. **Output Parser:** Die Antwort der KI (oft JSON) wird geparst und validiert.
6. **Action Node:** Das Ergebnis wird verarbeitet (Jira Ticket erstellen, E-Mail senden, Datei speichern).

Visuelle Metapher für den Bericht: Stellen Sie sich die Workflow Automation wie eine digitale Fertigungsstraße vor. Die Rohstoffe (Daten) kommen rein, Roboterarme (Nodes) bearbeiten sie. Aber an einer Station sitzt jetzt kein starrer Roboter mehr, sondern ein intelligenter Handwerker (das KI-Modell), der Entscheidungen trifft, Qualitätskontrollen durchführt und improvisieren kann, wenn das Werkstück leicht von der Norm abweicht. Das ist der Unterschied zwischen klassischer Automatisierung und KI-Agenten.

Abbildung 1: Quelle: https://razzfazz.ai/

6. Strategische Einordnung: Cloud vs. Local im Kostenvergleich

Ein häufiges Gegenargument gegen lokale KI sind die Anschaffungskosten (CAPEX). „Die Cloud ist doch billiger, ich zahle nur, was ich nutze." Das ist ein Trugschluss, besonders bei „Reasoning"-Modellen und agentischen Workflows.

1. **Token-Ökonomie:** Reasoning-Modelle wie Magistral oder OpenAI o1 generieren massive Mengen an „internen Gedanken-Tokens". In der Cloud bezahlen Sie für jeden dieser Gedanken. Ein einziger komplexer Refactoring-Task kann tausende Input- und zehntausende Output-Tokens verbrauchen. Bei täglicher Nutzung durch ein Entwicklerteam summieren sich diese OPEX (Operational Expenditures) enorm.
2. **Flatrate-Effekt:** Die razzfazz.ai Box ist eine Einmalinvestition. Ob Sie das Modell 10-mal oder 10.000-mal am Tag fragen, kostet dasselbe (abgesehen vom Strom). Das fördert Experimentierfreude. Entwickler zögern nicht, die KI zu nutzen, weil „das Budget knapp ist".
3. **Latenz:** Für Agenten, die viele kleine Schritte hintereinander ausführen (Schleifen), addiert sich die Netzwerklatenz der Cloud. Lokal erfolgt der Aufruf im Millisekundenbereich über localhost.

7. Fazit: Local matters.

Der Titel dieses Artikels ist Programm. „Local matters" ist nicht nur eine technologische Aussage, es ist eine Haltung. Der EU AI Act hat Risikoklassen definiert und fordert Transparenz und Governance. Viele Unternehmen sehen das als Bürde. Wir bei SEQIS sehen es als Chance.

Lösungen wie die **razzfazz.ai Box** beweisen, dass Compliance kein Innovationshemmer sein muss. Im Gegenteil: Indem wir die KI lokal betreiben, gewinnen wir Freiheiten zurück, die wir in der Cloud längst aufgegeben hatten.

- Wir können **Magistral** nutzen, um unsere geheimsten Geschäftsstrategien zu analysieren.
- Wir können **Devstral** auf unseren wertvollsten Asset – unseren Source Code – loslassen.
- Wir können Systeme bauen, die tief in unsere IT-Landschaft integriert sind, ohne Firewalls zu durchlöchern.

Die Zukunft der KI ist hybrid, aber das Herzstück – die Verarbeitung sensibler Daten und intellektuellen Eigentums – muss lokal liegen. Unified Memory Hardware, Open-Source-Modelle und intelligente Orchestrierung mittels der Workflow Automation machen dies heute möglich.

Ich lade Sie ein: Warten Sie nicht darauf, dass die Cloud sicher wird. Holen Sie sich die Sicherheit ins Haus. Experimentieren Sie. Bauen Sie Ihren ersten lokalen Agenten. Lassen Sie uns gemeinsam die Zukunft der Softwarequalität gestalten – effizient, souverän und lokal.

Ihr Alexander Vukovic
SEQIS Founder & Chief Evangelist

www.razzfazz.ai

Die folgende Tabelle illustriert diesen Effekt:

Kostenfaktor	Cloud AI (z.B. GPT-4 API)	Local AI (razzfazz.ai Box)
Reasoning (Denkzeit)	Teuer (wird per Token abgerechnet)	Kostenlos (inkludiert)
Code-Scan (großer Kontext)	Sehr teuer (großer Input-Kontext)	Kostenlos
Wiederkehrende Tasks	Linear steigende Kosten	Fixkosten (Amortisation)
Datentransfer	Kosten + Sicherheitsrisiko	Kein Transfer, keine Kosten

Quellen und weiterführende Informationen:

MIT-SEQIS QualityNews 2025-2_ AI Act & Local AI – Wege in eine verantwortungsvolle Zukunft-181125-221208.pdf

SEQIS - Ihre lokale AI Beratung, Zugriff am November 18, 2025, https://seqis.ai/

Razzfazz - Lokal. Open Source. Unabhängig. Sicher., Zugriff am November 18, 2025, https://razzfazz.ai/

Magistral - Mistral AI, Zugriff am November 18, 2025, https://mistral.ai/news/magistral

Devstral - Mistral AI, Zugriff am November 18, 2025, https://mistral.ai/news/devstral

Tutorial: Offline Agentic coding with llama-server · ggml-org llama.cpp · Discussion #14758, Zugriff am November 18, 2025, https://github.com/ggml-org/llama.cpp/discussions/14758

Magistral: The Open-Source AI That Thinks Like You | by Mayank Sultania | Medium, Zugriff am November 18, 2025, https://medium.com/@mayanksultania/magistral-the-open-source-ai-that-thinks-like-you-72af41d5cb64

mistralai/Magistral-Small-2509 - Hugging Face, Zugriff am November 18, 2025, https://huggingface.co/mistralai/Magistral-Small-2509

magistral - Ollama, Zugriff am November 18, 2025, https://ollama.com/library/magistral

Magistral: Mistral's Open Source Reasoning Model - Apidog, Zugriff am November 18, 2025, https://apidog.com/blog/magistral/

mistralai/Devstral-Small-2507 - Hugging Face, Zugriff am November 18, 2025, https://huggingface.co/mistralai/Devstral-Small-2507

Devstral Small 1.0 - Open Laboratory, Zugriff am November 18, 2025, https://openlaboratory.ai/models/devstral-small-2505

Exploring Reasoning LLMs and Their Real-World Applications - GetStream.io, Zugriff am November 18, 2025, https://getstream.io/blog/reasoning-llms/

Agentic Code Generation Papers Part 1, Zugriff am November 18, 2025, https://cbarkinozer.medium.com/agentic-code-generation-papers-part-1-05546d0d5d23

Devstral Small: The best Software Engineering Agentic LLM by Mistral | by Mehul Gupta | Data Science in Your Pocket | Medium, Zugriff am November 18, 2025, https://medium.com/data-science-in-your-pocket/devstral-small-the-best-software-engineering-agentic-llm-by-mistral-4b47b72ae705

mistralai/Devstral-Small-2507 : r/LocalLLaMA - Reddit, Zugriff am November 18, 2025, https://www.reddit.com/r/LocalLLaMA/comments/1lwe5y8/mistralaidevstralsmall2507/

What Is a Reasoning Model? | IBM, Zugriff am November 18, 2025, https://www.ibm.com/think/topics/reasoning-model

Alexander Vukovic ist SEQIS Gründer und Chief Evangelist.

Er ist erster Ansprechpartner für alle agilen, testmethodischen und testtechnischen Anfragen. In der Praxis arbeitet er als Agile Quality Coach, Berater, Interims-Testmanager, CI-Experte und Lasttester. Mehr als 25 Jahre Beratertätigkeit führten ihn während seiner zahlreichen Projekte in die unterschiedlichsten Branchen und Länder.

Sein persönliches Motto „Es gibt keine Probleme, sondern nur nicht gefunden Lösungen" spiegelt sich in jedem Projekt wider.

razzfazz.ai - Die Box für Ihre Local AI

Die razzfazz.ai Box bringt Enterprise-KI zurück ins eigene Haus. Durch lokale Inferenz entfallen Datenschutzrisiken und laufende Token-Kosten.

Hardware-Power: Dank Unified Memory teilen sich CPU und GPU bis zu 128 GB Arbeitsspeicher. Das ermöglicht den Betrieb riesiger KI-Modelle (bis 70B Parameter), die auf normalen Grafikkarten keinen Platz finden – und das flüsterleise bei nur 30–300 Watt.

Spezialisierte Intelligenz: Optimiert für Magistral (für logische Analysen) und Devstral (für Coding), bietet die Box leistungsstarke KI, die komplexe Aufgaben wie Test-Design oder Refactoring autonom löst.

Voll integriert: Der Stack aus llama.cpp (Inferenz), Workflowautomationen und Qdrant (Vektordatenbank) macht die Box zur Schaltzentrale für Ihre Daten – alles lokal, sicher und DSGVO-konform.

Wirtschaftlich: Einmal investieren (CAPEX), unbegrenzt nutzen. Keine Kosten pro Token, keine Netzwerklatenz, volle Datensouveränität.

www.razzfazz.ai

Der AI Act im Jahr 2025 - Was bisher geschah...

von Alexander Lewisch

Abbildung 1: Quelle: Bild generiert von Gemini (Google AI)

Der EU AI Act 2025: Analyse der Implementierung, Kontroversen und globalen Positionierung

Einleitung: Vom Legislativakt zur gelebten Realität

Der im Mai 2024 vom Rat der Europäischen Union verabschiedete EU AI Act wurde als das weltweit erste umfassende Regelwerk zur Regulierung künstlicher Intelligenz (KI) positioniert. Basierend auf einem risikobasierten Ansatz, der KI-Systeme in vier Kategorien (von unannehmbarem Risiko bis minimalem Risiko) einteilt, verfolgt die Verordnung das duale Ziel, die Grundrechte der EU-Bürger zu schützen und gleichzeitig einen harmonisierten Rechtsrahmen für Innovation zu schaffen. Als Verordnung konzipiert, besitzt sie unmittelbare Geltung in allen Mitgliedstaaten, ohne dass eine nationale Umsetzung erforderlich wäre.

Der nachfolgende Beitrag versucht nun die Transformation dieser legislativen Theorie in die operative Praxis im Jahr 2025 zu analysieren. Während der AI Act am 1. August 2024 formell in Kraft trat, war das Jahr 2025 durch die ersten Phasen der gestaffelten Rechtswirksamkeit, die Etablierung einer neuen Governance-Architektur und, ganz entscheidend, durch erhebliche Implementierungshürden, eine Fragmentierung der Aufsicht und intensive politische Kontroversen geprägt. Daher kann 2025 schon jetzt als das Jahr gesehen werden, in dem die praktischen und politischen Durchsetzungsprobleme beginnen, die legislative Errungenschaft in Frage zu stellen.

Die Etablierung der Governance-Architektur

Die Governance-Bestimmungen der Verordnung in Kapitel VII traten am 2. August 2025 formell in Kraft. Dies operationalisierte die dreigliedrige Aufsichtsstruktur, bestehend aus dem Europäischen AI Office, dem AI Board und einem wissenschaftlichen Gremium unabhängiger Sachverständiger.

Das EU AI Office

Das EU AI Office wurde bereits durch einen Kommissionsbeschluss am 21. Februar 2024 eingerichtet und nahm am 2. August 2025 seine volle operative Tätigkeit auf. Es ist als integraler Bestandteil der Europäischen Kommission (Generaldirektion Kommunikationsnetze, Inhalte und Technologien) strukturiert. Die Kernzuständigkeit des AI Office liegt in der Überwachung und Durchsetzung der neuartigen Regeln für GPAI-Modelle sowie in der Sicherstellung einer kohärenten Anwendung der Verordnung in der gesamten Union, oft in Zusammenarbeit mit nationalen Behörden.[1] Bereits im Jahr 2025 war das AI Office hochaktiv. Es war federführend bei der Finalisierung des kontroversen „Code of Practice" für GPAI-Modelle und veröffentlichte erste Leitlinien, etwa zur Auslegung der Verbote nach Artikel 5 (in Kraft seit 2. Februar 2025) sowie Meldevorlagen für schwerwiegende Vorfälle im Zusammenhang mit GPAI-Modellen (gültig ab November 2025).[2]

Das AI Board und das wissenschaftliche Gremium

Parallel zum AI Office nahm das AI Board am 2. August 2025 seine Arbeit auf. Dieses Gremium besteht aus hochrangigen Vertretern der Mitgliedstaaten und hat eine beratende sowie koordinierende Funktion. Es soll die Kommission und die Mitgliedstaaten bei einer „konsistenten und pragmatischen Anwendung" des AI Act unterstützen. Als dritte Säule wurde die Einrichtung eines wissenschaftlichen Gremiums unabhängiger Sachverständiger im Sommer 2025 eingeleitet. Dessen definierte Aufgabe ist die wissenschaftlich-technische Beratung des AI Office, insbesondere bei der Bewertung der systemischen Risiken von GPAI-Modellen. Die Etablierung dieses Gremiums verlief jedoch schleppend. Berichten zufolge war die Arbeit des Gremiums Verzögerungen unterworfen, und ein Aufruf zur Einreichung von Bewerbungen für Experten lief bis Mitte September 2025. Dies indiziert, dass das Gremium Ende 2025 noch nicht voll funktionsfähig war.[3]

Die Governance-Struktur des AI Act startete 2025 somit mit einem signifikanten operativen Ungleichgewicht. Das zentralisierte, exekutive Organ war voll handlungsfähig und trieb die Agenda, insbesondere bei der GPAI-Regulierung, voran. Demgegenüber hinkten die als Korrektiv vorgesehenen Instanzen massiv hinterher: das unabhängige wissenschaftliche Gremium, dessen Expertise gerade für die GPAI-Bewertung essenziell ist, war verspätet, während die dezentralen Aufsichtsbehörden in den Mitgliedstaaten, zuständig für die

Durchsetzung vor Ort, in Schlüsselländern wie Deutschland nicht einmal benannt waren. Dieses Machtvakuum konzentrierte die operative Autorität in der kritischen Anfangsphase der Implementierung stark bei der Europäischen Kommission.[4]

Abbildung 2: Quelle: Bild generiert von Gemini (Google AI)

Die gestaffelte Rechtswirksamkeit - Analyse der Implementierungsphasen 2025

2. Februar 2025 - Die Verbote
Die erste Welle der Rechtswirksamkeit trat am 2. Februar 2025 in Kraft. Seit diesem Datum gelten die Verbote für KI-Praktiken mit unannehmbarem Risiko (Kapitel II, Artikel 5). Von entscheidender Bedeutung ist, dass diese Verbote auch für Systeme gelten, die bereits vor diesem Datum in Verkehr gebracht wurden. Unternehmen waren somit gezwungen, ihre bestehenden Systeme einer sofortigen Re-Evaluierung zu unterziehen.[5] Ebenfalls seit Februar 2025 gelten die Verpflichtungen zur „AI Literacy" (KI-Kompetenz) nach Artikel 4, die Anbieter und Betreiber zu Schulungsmaßnahmen verpflichten.

2. Mai 2025 - Verpasste Frist: Der GPAI-Kodex
Gemäß Artikel 56(9) des AI Acts sollte der „Code of Practice" für GPAI-Modelle bis zum 2. Mai 2025 fertiggestellt sein.[6] Diese Frist wurde aufgrund der Komplexität und der kontroversen Verhandlungen mit der Industrie verpasst. Der finale Kodex wurde erst im Juli 2025 veröffentlicht.[7]

2. August 2025 - GPAI und Governance
Die zweite Welle der Rechtswirksamkeit trat am 2. August 2025 in Kraft. Seit diesem Datum gelten die Verpflichtungen für Anbieter von GPAI-Modellen (Kapitel V). Gleichzeitig wurden die Governance-Bestimmungen (Kapitel VII) sowie die Bestimmungen zu Vertraulichkeit und Sanktionen (Kapitel XII) rechtswirksam. Die Sanktionen sind drakonisch und reichen bis zu 35 Millionen Euro oder 7% des weltweiten Jahresumsatzes für Verstöße gegen die Verbote des Artikel 5.

Die gestaffelte Implementierung erzeugte 2025 einen „Compliance-Rückstau", der durch eine Asynchronität von Rechtspflicht und Rechtsdurchsetzung gekennzeichnet ist. Seit Februar 2025 ist die Nutzung eines verbotenen Systems (Art. 5) rechtswidrig und kann theoretisch bestraft werden. Gleichzeitig, wie noch später ausgeführt wird, hatten zentrale Mitgliedstaaten wie Deutschland im August 2025 die für die Durchsetzung zuständige Aufsichtsbehörde noch nicht einmal benannt. Diese Fälle der Rechtswirksamkeit ohne Exekutive schuf eine Periode der Rechtsunsicherheit, die den Nährboden für Lobbying gegen die Bestimmungen des AI Acts bildete.[8]

Die Verbote des Artikel 5 in der Praxis
Das Inkrafttreten der Verbote für KI mit inakzeptablem Risiko am 2. Februar 2025 markierte den ersten Praxistest der Verordnung. Die unmittelbare Veröffentlichung von Kommissionsleitlinien zur Auslegung dieser Verbote nur zwei Tage später (4. Februar 2025) unterstreicht den hohen Grad an rechtlicher Unsicherheit, der diese Bestimmungen umgibt. Entgegen der öffentlichen Wahrnehmung sind die Verbote des Artikel 5 juristisch „weich", da sie durch unbestimmte Rechtsbegriffe und weitreichende Ausnahmen gekennzeichnet sind.[9]

Zivilgesellschaftliche Organisationen wie AlgorithmWatch kritisierten die Entwürfe der Kommissionsleitlinien scharf. Sie warnten, das Gesetz sei voller schwerwiegender Schlupflöcher. Die Kernkritik lautet, dass bestimmte Verbote, wie z.B. bei Predictive Policing, präzisiert und erweitert werden müssten und dass die Regulierung sich nicht an technischer Komplexität, sondern an potenziellen Folgeschäden orientieren müsse.[10]

Das Inkrafttreten der Verbote war somit nicht das Ende dieser Praktiken, sondern der Beginn einer intensiven juristischen Auseinandersetzung um deren Auslegung. Während Gerichtsentscheidungen Ende 2025 noch ausstanden, zeigten sich erste Tendenzen bei der behördlichen Durchsetzung, da nationale Datenschutzbehörden (DPAs) begannen die Verzögerungen nationaler Umsetzung und das daraus resultierende Governance-Vakuum zu nutzen, um die Zuständigkeiten an sich zu ziehen. In Spanien kündigte die Datenschutzbehörde (AEPD) an, dass sie auf Basis ihrer bestehenden Befugnisse aus der DSGVO gegen verbotene KI-Systeme nach Art. 5 vorgehen könne, die personenbezogene Daten verarbeiten, sogar noch bevor AEPD formell zur KI-Aufsichtsbehörde ernannt wird. Dies bedeutet, dass die Durchsetzung des AI Acts stark in Richtung Datenschutzrecht gelenkt wird.[11]

Das Epizentrum der Kontroverse 2025: Die Regulierung von GPAI
Während GPAI-Modelle bisher lediglich als eine von vier Risikostufen mit Transparenzpflichten eingeschätzt wurden, entwickelte sich dieses Thema 2025 zur zentralen Kontroverse über die Zukunft des AI Acts. Seit dem 2. August 2025 unterliegen Anbieter von GPAI-Modellen den Verpflichtungen nach Artikel 53, d.h. primär technische Dokumentation

und Einhaltung des EU-Urheberrechts. Anbieter von GPAI-Modellen mit systemischem Risiko unterliegen zusätzlich den verschärften Pflichten nach Artikel 55, d.h. Bewertung systemischer Risiken, rigorose Tests und Cybersicherheit.

Um diese abstrakten gesetzlichen Pflichten zu konkretisieren, sieht Artikel 56 einen freiwilligen Verhaltenskodex (CoP - Code of Practice) vor, der sich zunehmend zu einem Konfliktherd entwickelt hat. Dessen Einhaltung gewährt Anbietern eine Konformitätsvermutung (presumption of compliance), was den Verwaltungsaufwand und die Rechtsunsicherheit massiv reduziert.[12] Dieser von Experten entwickelte Kodex verpasste seine ursprüngliche Frist (2. Mai 2025) und wurde erst am 10. Juli 2025 vom AI Office veröffentlicht und als adäquat bestätigt.[13] Unmittelbar nach der Veröffentlichung kam es zum Eklat, der die „Big Tech"-Industrie spaltete. Während Google (Alphabet) am 30. Juli 2025 ankündigte, den Kodex unterzeichnen zu wollen[14], und Microsoft den Verhaltenskodex ausdrücklich befürwortet hat15, verweigerte Meta (Facebook) die Unterzeichnung des CoP[16]. Die Zustimmung Googles erfolgte jedoch unter scharfem Vorbehalt, da Vertreter von Google öffentlich die Sorge äußerten, der AI Act und der Kodex würden Europas Entwicklung und Einsatz von AI verlangsamen sowie die Wettbewerbsfähigkeit gefährden.[17] Meta bezeichnete den Kodex als zweideutig, exzessiv und als Bedrohung für das Wachstum. Meta betonte, man werde sich zwar an den AI Act halten, jedoch nicht an den freiwilligen Kodex.[18]

Metas Weigerung stellt eine direkte strategische Herausforderung an den gesamten Durchsetzungsmechanismus der Kommission dar. Der Plan des AI Office basierte darauf, den CoP als skalierbaren Compliance-Pfad zu etablieren. Meta entkoppelt nun den freiwilligen Kodex vom mandatorischen Gesetz und zwingt das AI Office, die Konformität von Metas Modellen direkt anhand des abstrakten Gesetzestextes von Art. 53/55 zu prüfen. Dies ist für die neue Behörde ein ressourcenintensiver und juristisch hochriskanter Weg.[19] Diese Auseinandersetzung politisierte die gesamte Implementierung. Die technische Kritik am CoP, die auch von Industrieverbänden wie der CCIA[20] und europäischen Konzernen wie Siemens, SAP und Airbus geteilt wurde[21], mündete in der politischen Forderung nach einem Stopp der Verordnung oder zumindest einer Regulierungspause („potential EU AI Act pause").[22]

Nationale Umsetzungsstrategien im Vergleich

Obwohl der AI Act als Verordnung konzipiert ist, die unmittelbar gilt und Harmonisierung zum Ziel hat, erfordert sie von den Mitgliedstaaten die Benennung zuständiger nationaler Behörden und Marktüberwachungsbehörden. Die Frist hierfür war der 2. August 2025. In der Praxis führte dieser Prozess 2025 zu einem paradoxen Ergebnis. Statt Harmonisierung entstand eine tiefgreifende regulatorische Fragmentierung. Bereits im November 2024 hatten nur 3 von 27 Mitgliedstaaten ihre zuständigen Behörden klar benannt.[23]

Österreich

Österreichs Umsetzungsstrategie im Jahr 2025 war von Verzögerungen und einem Mangel an regulatorischer Klarheit geprägt. Obwohl das Land proaktiv eine „KI-Servicestelle" (AI Service Desk) bei der Rundfunk und Telekom Regulierungs-GmbH (RTR-GmbH) eingerichtet hat, dient diese primär als Informations- und Anlaufstelle für Unternehmen. Entscheidend ist, dass Österreich die Frist vom 2. August 2025 zur Benennung der zentralen Marktüberwachungsbehörde und der notifizierenden Behörde nicht eingehalten hat. Ein entsprechendes nationales Durchführungsgesetz, das ursprünglich für das erste Quartal 2025 erwartet wurde, ist bisher auch noch nicht verabschiedet. Zwar wurde im November ein „KI-Umsetzungsplan" sowie eine Liste von 19 Stellen zum Schutz der Grundrechte (gemäß Artikel 77) veröffentlicht, die Kernzuständigkeit für die Marktaufsicht blieb jedoch vakant.[24] Diese regulatorische Unsicherheit trifft auf eine Wirtschaft, die laut Studien (z.B. McKinsey „State of AI in Austria 2025") bei der AI-Reife (AIQ-Score 30) hinter dem EU-Durchschnitt (34) zurückliegt. Lediglich 20 % der österreichischen Unternehmen gaben an, eine ausformulierte KI-Strategie zu besitzen.[25]

Deutschland

Deutschland hat die Frist vom 2. August 2025 zur Benennung seiner Aufsichtsbehörde verpasst. Dieser Zustand führte zu massiver Rechtsunsicherheit bei Unternehmen und scharfer Kritik von Datenschützern, die vor einer Kontrolllücke und fehlenden Ansprechpartnern warnten. Die politische Debatte in Deutschland im Rahmen des KI-Marktüberwachungs-Gesetzes ist aktuell noch nicht abgeschlossen und dreht sich um die Zuständigkeit: die Bundesnetzagentur, die Datenschutzbehörden oder eine neu zu gründende Behörde.[26]

Italien

Italien verfolgte einen diametral entgegengesetzten, proaktiven Ansatz. Die Regierung erließ ein umfassendes eigenes nationales KI-Gesetz (Legge Nr. 132/2025), das am 10. Oktober 2025 in Kraft trat.[27] Dieses Gesetz ergänzt den AI Act und führt zusätzliche nationale Regeln („Gold-Plating") ein. Dazu gehören erweiterte Transparenzpflichten für Arbeitgeber beim Einsatz von KI am Arbeitsplatz und ein neuer Straftatbestand für die unrechtmäßige Verbreitung von KI-generierten Inhalten (Deepfakes), der mit ein bis fünf Jahren Haft bestraft wird. Die Benennung der Aufsichtsbehörden wurde durch das Gesetz an die Regierung delegiert.[28]

Frankreich und Spanien
Frankreich kündigte im September 2025 einen nationalen Aufsichtsplan an, der auf eine Aufteilung der Zuständigkeiten zwischen bestehenden Regulierungsbehörden wie der Datenschutzbehörde, der Wettbewerbsbehörde und der Medienaufsicht hinauszulaufen scheint.[29] Spanien, das die Einrichtung einer dedizierten KI-Behörde (AESIA) plant, sah sich mit einer proaktiven Zuständigkeitserklärung seiner Datenschutzbehörde (AEPD) konfrontiert.[30]

Diese Entwicklungen untergraben das Kernziel eines harmonisierten Binnenmarktes. Ein Anbieter von AI, der 2025 in der EU tätig ist, sieht sich einem regulatorischen Vakuum in Deutschland, zusätzlichen Strafgesetzen in Italien und einem sektoralen Patchwork in Frankreich gegenüber.

Der "Brussels Effect" auf dem Prüfstand
Die Hoffnung, der AI Act werde einen „Brussels Effect" auslösen und de facto zum globalen Standard werden, bleibt im Jahr 2025 stark umstritten. So insistiert etwa der Politikwissenschaftler Ben Crum, der AI Act sei kein klassischer „Brussels Effect". Er geht davon aus, dass AI, anders als frühere Regulierungsthemen wie z.B. der Datenschutz, von fundamentaler Unsicherheit und existenziellem Risiko geprägt sei. Crum schlägt vor, den AI Act nicht als globalen Standard, sondern als experimentelle Regulierung zu verstehen, ein regionaler Ansatz unter vielen, der sich kooperativ weiterentwickeln müsste.[31]

Die EU befindet sich 2025 in einer geopolitischen Zwickmühle, da sie mit ihrem Weg der menschenzentrierten Regulierung isoliert ist. Während die USA auf aktive Deregulierung setzen und unter Präsident Trump bereits erste AI-Sicherheitsauflagen wieder gekippt wurden, forciert China die staatliche Technologieentwicklung.[32] Wenn US-Unternehmen wie Meta beginnen, EU-Compliance-Mechanismen wie den Verhaltenskodex offen abzulehnen, und die US-Regierung dies durch Deregulierung politisch deckt, erodiert die globale Normsetzungsmacht der EU. Aus Sicht der USA bleibt das bisherige Paradigma „the U.S. innovates, and the EU regulates" auch im Jahr 2025 weitgehend bestehen.[33]

Abbildung 3: Quelle: Bild generiert von Gemini (Google AI)

Der AI Act im Alltag 2025: Erste sichtbare Implikationen
Im Alltag der Bürger und Unternehmen im Jahr 2025 sind die Auswirkungen des AI Act zweigeteilt, in Transparenz für Bürger gem. Artikel 50 und Pflichten für Unternehmen gem. Artikel 4 & 5. Der AI Act agiert somit gleichzeitig als Verbraucherschutz-Informationsgesetz und als Teil des Arbeits- und Organisationsrechts.

Für Bürger als Konsumenten von Medien sind die „weichen" Transparenzpflichten des Artikel 50 am sichtbarsten. Bei der Nutzung von Chatbots müssen Nutzer informiert werden, dass sie mit einem KI-System interagieren. Künstlich erzeugte oder manipulierte Audio-, Bild- oder Videoinhalte müssen als solche gekennzeichnet werden.[34] Angesichts globaler Sorgen vor Desinformation und neuer nationaler Gesetze treibt die EU-Kommission dieses Thema aktiv voran und hat am 5. November 2025 die Arbeit an einem separaten „Verhaltenskodex zur Kennzeichnung von KI-generierten Inhalten" aufgenommen.[35]

Für Unternehmen als Betreiber und Arbeitgeber ist die Wirkung hart und obligatorisch. Die, laut Artikel 4 (AI Literacy), seit Februar 2025 geltende Pflicht zur Sicherstellung von „AI-Kompetenz" bei Personal, das KI-Systeme (insb. Hochrisiko-Systeme) betreibt, schafft einen neuen Compliance- und Bildungssektor. Unternehmen sind verpflichtet in Schulungen zu investieren und Governance-Strukturen aufzubauen. Trotz dieser gesetzlichen Verpflichtung zum Aufbau der AI-Kompetenz in Unternehmen, zeigt eine neue repräsentative Umfrage des Digitalverbands Bitkom, dass in Deutschland nur 20% der Beschäftigten eine AI-Schulung erhalten und sogar 70% der Firmen die AI-Schulungspflicht gänzlich ignorieren.[36]

Für Unternehmen kommt auch das Verbot hinzu, Emotionserkennung am Arbeitsplatz und in Bildungseinrichtungen einzusetzen (Artikel 5 - Verbot der Emotionserkennung), das seit Februar 2025 rechtsverbindlich ist und den Einsatz entsprechender HR- oder Bildungssoftware illegal macht.

Conclusio und Ausblick: Die Zukunftssicherheit des AI Acts

Das Jahr 2025 markierte den Übergang des AI Acts von einer legislativen Ambition in eine operative Realität. Der 2024 beschriebene Meilenstein wurde 2025 durch vier Schlüsselfaktoren auf die Probe gestellt. Die (1) Etablierung einer zentralen Governance (AI Office), wird durch die (2) dezentrale Fragmentierung durch nationale Verzögerungen konterkariert. Das (3) Inkrafttreten erster Rechtsfolgen ist nahezu unwirksam, weil deren Durchsetzung mangels Behörden unsicher ist. Hinzu kommt der (4) strategische Konflikt mit globalen GPAI-Anbietern, der den Zeitplan und den Kern der Regulierung bedroht und in Frage stellt.

Daher plant die Europäische Kommission zentrale Teile ihres historischen Gesetzes zur Regulierung künstlicher Intelligenz bereits jetzt abzuschwächen bzw. deren Einführung zu verzögern. Dies geschieht als Reaktion auf massiven Druck von US-Technologiekonzernen und der US-Regierung. So erwägt die EU eine Schonfrist für Unternehmen, das Aussetzen von Geldstrafen bis August 2027 sowie bestimmte Vereinfachungen und Ausnahmen. Über diese Maßnahmen soll noch im November 2025 entschieden werden, um die Wettbewerbsfähigkeit der EU gegenüber den USA und China zu stärken und Bedenken der Industrie über zu hohe Compliance-Kosten und Innovationshemmnisse auszuräumen. Während Wirtschaftsvertreter die Pläne als notwendigen Schritt begrüßen, um Innovationen in Europa nicht abzuwürgen, sind Bürgerrechts- und Datenschutzorganisationen alarmiert, weil sie davor warnen, dass diese Änderungen die Sicherheitsstandards aushöhlen und den Schutz der Bürger schwächen könnten.[37]

Der AI Act droht also operativ zu scheitern, noch bevor seine wichtigsten Bestimmungen für Hochrisiko-KI-Systeme im Jahr 2026 überhaupt in Kraft treten. Die Implementierung der vermeintlich einfacheren Governance- und GPAI-Regeln im Jahr 2025 sollte die Vorbereitungsphase sein. Diese ist in wesentlichen Teilen misslungen. Wenn die EU bereits an der Etablierung der Struktur scheitert, ist ernsthaft in Zweifel zu ziehen, wie sie 2026 die kommenden Inhalte durchsetzen will. Die politische und operative Basis der Verordnung erodierte 2025 schneller als ihr eigener gestaffelter Zeitplan fortschreitet.

Downloads

McKinsey & Company: State of AI in Austria 2025. Wie Unternehmen ihre Produktivität und Wettbewerbsfähigkeit steigern. https://www.mckinsey.com/de/~/media/mckinsey/locations/europe%20and%20middle%20east/austria/2025-09%20ki-reife%20oesterreich/mckinsey_ki%20in%20oesterreichs%20unternehmen_september%202025.pdf

Quellen

[1] CMS: Durchsetzung der KI-VO auf europäischer Ebene: EU AI Office. In: https://www.cmshs-bloggt.de/rechtsthemen/kuenstliche-intelligenz/durchsetzung-der-ki-vo-auf-europaeischer-ebene-das-eu-ai-office/

[2,9] Peters, Salome: EU-Kommission veröffentlicht Leitlinien zu verbotenen KI-Systemen. In: https://www.vdma.eu/viewer/-/v2article/render/140182315

[3,4] Amman, Thorsten; Achnitz, Florian; Tobey, Danny; Stokes Gareth; Dauzier, Jeanne; Darling, Coran & Blackford, Liam: Latest wave of obligations under the EU AI Act take effect: Key considerations. In: https://www.dlapiper.com/en-us/insights/publications/2025/08/latest-wave-of-obligations-under-the-eu-ai-act-take-effect

[5] Bitkom e.V.: Umsetzungsleitfaden zur KI-Verordnung. In: https://www.bitkom.org/sites/main/files/2024-10/241028-bitkom-umsetzungsleitfaden-ki.pdf

[6] EU Artificial Intelligence Act: Zeitplan für die Umsetzung. In: https://artificialintelligenceact.eu/de/implementation-timeline/

[7,12,22] IAPP: Navigate 2025: Potential EU AI Act pause opens new questions on approach to global regulation. In: https://iapp.org/news/a/navigate-2025-potential-eu-ai-act-pause-opens-new-questions-on-approach-to-global-regulation

[8,26] Grensemann, René: Deutschland verpasst AI-Act-Frist: Keine Aufsichtsbehörde benannt. In: https://www.aiact-akademie.de/aktuelles/deutschland-verpasst-ai-act-frist-keine-aufsichtsbehoerde

[10] AlgorithmWatch: Leitlinien zum „AI Act": Kritik aus der Zivilgesellschaft. In: https://algorithmwatch.org/de/leitlinien-zum-ai-act-kritik-aus-der-zivilgesellschaft/#

[11,30] One Trust Data Guidance: Spain: AEPD announces that it can act against prohibited AI systems that process personal data. In: https://www.dataguidance.com/news/spain-aepd-announces-it-can-act-against-prohibited-ai

[13] Europäische Kommission: Der General-Purpose AI Code of Practice (KI-Verhaltenskodex für allgemeine Zwecke). In: https://digital-strategy.ec.europa.eu/de/policies/contents-code-gpai

[14,17] Google: We will sign the EU AI Code of Practice. In: https://blog.google/around-the-globe/google-europe/eu-ai-code-practice/

[15] Grensemann, René: Microsoft an der Spitze. Tech-Riese befürwortet EU-Verhaltenskodex für KI. In: https://www.aiact-akademie.de/aktuelles/microsoft-unterstuetzt-ai-pakt

[16] Grensemann, René: Google biegt auf EU-Kurs ab, Meta wählt die Konfrontation. In: https://www.aiact-akademie.de/aktuelles/google-unterzeichnet-ai-act

[18,19] Lundy, Jim: Meta vs. Brussels: Why the Social Media Giant Said 'No' to the EU AI Pact. In: https://aragonresearch.com/meta-said-no-to-eu-ai-pact/

[20] CCIA: AI Act: EU's Final GPAI Code Imposes Disproportionate Burden, Improvements Required. In: https://ccianet.org/news/2025/07/ai-act-eus-final-gpai-code-imposes-disproportionate-burden-improvements-required/

[21] Grensemann, René: AI Act: Deutschlands Zerreißprobe. Zwischen Fristchaos und Industrierevolte. In: https://www.aiact-akademie.de/aktuelles/ai-act-deutsche-zerreissprobe-regulierung-vs-industrie

[23,24] EU Artificial Intelligence Act: Überblick über alle nationalen Umsetzungspläne des AI-Gesetzes. In: https://artificialintelligenceact.eu/de/national-implementation-plans/

[25] McKinsey & Company: State of AI in Austria 2025 Wie Unternehmen ihre Produktivität und Wettbewerbsfähigkeit steigern. September 2025. In: https://www.mckinsey.com/de/~/media/mckinsey/locations/europe and middle east/austria/2025-09 ki-reife oesterreich/mckinsey_ki in oesterreichs unternehmen_september 2025.pdf

[27] Barbiero, Adelaide; Campasso, Biancamaria; Cipriani, Augusto; Di Cillo, Roberto & Scelta, Daniele: Trust at the core: Italy's first comprehensive AI Law and its impact on the professional practice. In: Trust at the core: Italy's first comprehensive AI Law and its impact on the professional practice - MediaLaws

[28] Norton Rose Fulbright LLP: Italy enacts Law No. 132/2025 on Artificial Intelligence: Sector rules and next steps. In: https://www.nortonrosefulbright.com/en/knowledge/publications/9bfedfea/italy-enacts-law-no-132-2025-on-artificial-intelligence-sector-rules-and-next-steps

[29] One Trust Data Guidance: France: Ministry announces national oversight plan for AI regulation. In: https://www.dataguidance.com/news/france-ministry-announces-national-oversight-plan-ai

[31] Crum, Ben: Brussels effect or experimentalism? The EU AI Act and global standard-setting. Internet Policy Review 14.3 (2025). In: https://policyreview.info/articles/analysis/brussels-effect-or-experimentalism

[32] DIN e. V.: AI Action Summit 2025: Europa im Zentrum der KI-Debatte. In: https://www.din.de/de/din-und-seine-partner/presse/mitteilungen/ai-action-summit-2025-1200068

[33] Kästle, Vincent M. & Wolfenstätter, Tobias: The US Innovates, the EU Regulates? Contrasting Approaches to AI Regulation Across the Atlantic. In: https://www.dajv.de/ai-act/the-us-innovates-the-eu-regulates-contrasting-approaches-to-ai-regulation-across-the-atlantic/

[34] Heynike, Francois: AI Act und generative KI: Was Unternehmen jetzt wissen sollten. In: AI Act und generative KI: Was Unternehmen jetzt wissen sollten

[35] Europäische Kommission: Europäischer Ansatz für künstliche Intelligenz. In: https://digital-strategy.ec.europa.eu/de/policies/european-approach-artificial-intelligence

[36] Grensemann, René: Die große KI-Kompetenzlücke. Bitkom-Studie: 70% der Firmen ignorieren Schulungspflicht. In: https://www.aiact-akademie.de/aktuelles/bitkom-umfrage-ki-schulungspflicht

[37] Moens, Barbara: EU set to water down landmark AI act after Big Tech pressure. Artikel Financial Times vom 7. November 2025. In: https://www.ft.com/content/af6c6dbe-ce63-47cc-8923-8bce-4007f6e1

Weiterführende Informationen

Zeitplan für die Umsetzung des AI Acts, siehe: https://artificialintelligenceact.eu/de/implementation-timeline/

Zeitplan als Bild, siehe https://artificialintelligenceact.eu/wp-content/uploads/2024/08/AIA-Implementation-Timeline-1-August-2024-v2.png

Alexander Lewisch ist Consultant bei SEQIS.

Als IT-Quereinsteiger und Wiedereinsteiger konnte er Erfahrungen im Software Testing sammeln und Einblicke in diverse Bereiche wie Testmanagement, Testdurchführung, Testautomation, Testkoordination, Reporting und Consulting sammeln.

Derzeit widmet er sich verschiedenen Testprojekten im Bereich der Remote Testing Services (RTS).

KI-Souveränität im Fokus: Wie Local AI die Datenhoheit sichert und den EU AI Act erfüllt

von Martin Wildbacher

Der Artikel beleuchtet die strategische Bedeutung von lokaler KI im Kontext des EU AI Acts und der Datensouveränität. Er vergleicht lokale mit Cloud-basierten KI-Systemen hinsichtlich Datenschutz, Kosten und Skalierbarkeit und zeigt auf, wie lokale KI die Einhaltung regulatorischer Anforderungen proaktiv unterstützt.

1. Einleitung: Das Spannungsfeld zwischen Innovation und Verantwortung

Die transformative Kraft der Künstlichen Intelligenz (KI) hat in den letzten Jahren nahezu alle Branchen revolutioniert. Von der Automatisierung komplexer Prozesse bis hin zur Generierung von Inhalten in Echtzeit ermöglichen KI-Systeme eine Effizienzsteigerung und Innovation, die noch vor wenigen Jahren undenkbar schien. Doch mit dem rasanten Fortschritt wächst auch eine grundlegende Sorge, die das Vertrauen der Nutzer und Unternehmen in diese Technologie auf eine harte Probe stellt: die Frage nach dem Schutz sensibler Daten und der Datenhoheit.

Eine Studie von KPMG aus dem Jahr 2024 ergab, dass 63% der Verbraucher Bedenken hinsichtlich der potenziellen Kompromittierung ihrer Privatsphäre durch generative KI haben, die persönliche Daten durch unbefugten Zugriff oder Missbrauch preisgeben könnte. Die Ängste reichen von der Überwachung durch Smart-Home-Geräte bis zur unbefugten Nutzung von Sprachdaten, was die Notwendigkeit von robusten Datenschutzlösungen unterstreicht. Die zunehmende „privacy resignation" – eine Resignation, bei der Nutzer akzeptieren, dass ihre Daten ohnehin geteilt werden – macht es für Unternehmen umso wichtiger, aktiv das Gegenteil zu beweisen und das Vertrauen durch den Schutz persönlicher Informationen zurückzugewinnen.[1] Genau hier setzt die Strategie der lokalen KI an. Anstatt die Datenhoheit aus den Händen zu geben, bietet sie einen Weg, die Innovationskraft der KI zu nutzen, ohne die Kontrolle über die Daten zu verlieren.

2. Grundlagen: Lokale vs. Cloud-basierte KI – Ein strategischer Vergleich

Die Wahl zwischen einem lokalen und einem Cloud-basierten KI-System ist eine der grundlegendsten strategischen Entscheidungen, die ein Unternehmen heute treffen muss. Beide Ansätze repräsentieren unterschiedliche Architekturen mit signifikanten Auswirkungen auf Datenverarbeitung, Sicherheit, Kosten und Flexibilität.

2.1 Die technischen und betriebswirtschaftlichen Architekturen

Cloud-basierte KI-Lösungen verarbeiten Daten auf externen Servern, die von Anbietern wie OpenAI oder Google betrieben werden.[2] Dieser Ansatz besticht durch seine hohe Skalierbarkeit, die Möglichkeit, mit großen Datenmengen umzugehen, und geringere anfängliche Investitionskosten, da die Rechenleistung lediglich gemietet wird.[2] Die globale Zugänglichkeit erleichtert zudem die Zusammenarbeit über verschiedene Standorte hinweg.[2]

Lokale KI-Systeme, oft auch als Edge AI bezeichnet, verarbeiten Daten hingegen direkt auf dem Endgerät oder im eigenen lokalen Netzwerk.[2] Die KI-Modelle laufen direkt auf der unternehmenseigenen Hardware, was die Latenzzeiten minimiert und eine Datenverarbeitung in Echtzeit ermöglicht.[2] Der Hauptvorteil dieses Ansatzes liegt in der Datenhoheit: Alle Daten bleiben innerhalb der unternehmenseigenen Infrastruktur, wodurch die Abhängigkeit von externen Cloud-Diensten reduziert wird und das Risiko von Datenlecks minimiert wird.[2]

2.2 Der Datenfluss im Detail

Der grundlegende Unterschied zwischen den beiden Architekturen lässt sich am besten anhand ihres Datenflusses visualisieren. Während bei der Cloud-KI Daten das lokale Netzwerk verlassen müssen, um zum externen Server des Anbieters gesendet zu werden, bleibt bei der lokalen KI der gesamte Verarbeitungsprozess intern.

Datenflussdiagramm: Lokale KI vs. Cloud KI

Ein Flussdiagramm für ein lokales KI-System würde den Datenweg als einen geschlossenen Kreislauf innerhalb des Unternehmensnetzwerks darstellen.[8] Der Prozess beginnt, wenn der Endnutzer Eingabedaten bereitstellt, die dann direkt von einem lokalen KI-Modell auf einem Gerät oder Server verarbeitet werden. Die Ausgabedaten werden, ohne das interne Netzwerk zu verlassen, an den Nutzer zurückgegeben. Der gesamte Datenpfad bleibt strikt innerhalb der Unternehmens-Firewall.[2]

Im Gegensatz dazu visualisiert der Datenfluss einer Cloud-KI eine externe Schleife.[8] Eingabedaten werden vom Endnutzer über das Internet an die externen Cloud-Server des Anbieters gesendet. Dort findet die Verarbeitung durch das KI-Modell statt, bevor die Ausgabedaten über das Internet an den Endnutzer zurückgeschickt werden.[2] Die externe Datenübertragung ist der kritische Punkt, der die Datenhoheit beeinträchtigt und potenzielle Sicherheitsrisiken mit sich bringt.[2]

Aspekt	Lokale KI	Cloud KI
Datenverarbeitung	Direkt auf dem Gerät/im lokalen Netzwerk[2]	Auf externen Servern des Anbieters[2]
Datenschutz & Souveränität	Maximale Kontrolle, Daten bleiben inhouse, DSGVO-konform[3]	Daten verlassen die lokale Umgebung, Abhängigkeit von Drittanbietern[2]
Latenz	Sehr gering, ideal für Echtzeitanwendungen[2]	Höher, da Datenübertragung erforderlich ist[2]
Kosten	Höhere Anfangsinvestitionen (Hardware, ggf. Lizenzen)[3]	Niedrigere Einstiegskosten, dafür laufende Gebühren[3]
Skalierbarkeit	Erfordert Hardware-Upgrades, weniger flexibel[3]	Dynamisch skalierbar durch Mieten zusätzlicher Kapazitäten[3]
Kontrolle & Unabhängigkeit	Volle Kontrolle über Modelle und Daten, unabhängig von externen Anbietern[2]	Abhängig vom Cloud-Anbieter, weniger Kontrolle über Modelle[2]

2.3 Tiefergehende Betrachtung der strategischen Nuancen

Die scheinbar offensichtlichen Vor- und Nachteile beider Systeme sind bei näherer Betrachtung differenzierter.

Das vermeintliche Kostendilemma ist oft eine Frage der Perspektive. Während Cloud-KI mit geringen Einstiegshürden lockt, können die laufenden, nutzungsbasierten Kosten bei wachsendem Datenvolumen und zunehmender Nutzung schnell anwachsen und langfristig „nicht nachhaltig" werden.[2] Eine höhere Anfangsinvestition in leistungsstarke, lokale Hardware kann dagegen über die Zeit zu niedrigeren Gesamtbetriebskosten (Total Cost of Ownership, TCO) führen, da keine wiederkehrenden Gebühren für Datentransfer oder Rechenleistung anfallen.[3]

Auch das gängige Missverständnis der Skalierbarkeit muss nuanciert betrachtet werden. Zwar sind die Kapazitäten einer lokalen Infrastruktur schneller erschöpft als die einer Cloud-Umgebung, die auf massive Rechenressourcen zurückgreifen kann.[3] Jedoch ermöglichen hybride Architekturen eine flexible Skalierung, indem sie die Vorteile beider Welten kombinieren.[3] Solche Lösungen behalten die datensensitiven Prozesse lokal, während sie für rechenintensive oder weniger kritische Aufgaben auf Cloud-Ressourcen zurückgreifen.
Ein Beispiel hierfür ist das Federated Learning, eine dezentrale Lernarchitektur, die im späteren Abschnitt genauer beleuchtet wird.

3. Local AI als Compliance-Motor: Die rechtliche Verknüpfung zum EU AI Act

Der EU AI Act ist das weltweit erste umfassende Gesetz zur Regulierung von KI. Er ist kein isoliertes Regelwerk, sondern eine Erweiterung und Konkretisierung bestehender Datenschutzprinzipien. Der Act stellt klar, dass die Datenschutz-Grundverordnung (DSGVO) in jedem Fall gilt, in dem personenbezogene Daten verarbeitet werden.[12] Er verankert zudem Schlüsselprinzipien wie Rechenschaftspflicht, Fairness und Transparenz, die bereits in der DSGVO festgeschrieben sind.[12]

3.1 Einhaltung der Kernprinzipien durch Local AI

Lokale KI-Lösungen bieten einen proaktiven Weg, die strengen Anforderungen des EU AI Acts zu erfüllen:

- Datenhoheit und Datenschutz by Design: Der AI Act fordert, dass Anbieter von KI-Systemen einen „Datenschutz by Design"-Ansatz verfolgen. Lokale KI erfüllt dieses Prinzip im Kern, da Daten nicht an externe Server übertragen

werden müssen.[3] Die Verarbeitung innerhalb des eigenen Netzwerks minimiert das Risiko von unbefugtem Zugriff und Datenlecks, was die Einhaltung der strengen DSGVO-Vorschriften erheblich erleichtert.[3]

- Rechenschaftspflicht (Accountability): Der AI Act macht Anbieter von Hochrisiko-KI-Systemen für die Einhaltung der Vorschriften verantwortlich und fordert eine systematische und geordnete Dokumentation der Compliance.[12] Durch die volle Kontrolle über Daten und Modelle innerhalb der eigenen Infrastruktur wird die Nachverfolgbarkeit (Traceability) der Verarbeitung vereinfacht und die Rechenschaftspflicht gestärkt.[2] Es existiert eine klare Verantwortungskette, was die Umsetzung der regulatorischen Anforderungen erleichtert.

- Transparenz: Die Vorschriften des AI Acts verlangen, dass Nutzer über die Interaktion mit einem KI-System informiert werden und KI-generierte Inhalte (wie Deepfakes) gekennzeichnet werden müssen.[13] Auch bei lokalen Systemen muss diese Transparenz gewährleistet sein, beispielsweise durch eine deutliche Benachrichtigung über die lokale Datenverarbeitung. Unternehmen, die sich für Local AI entscheiden, können dies als Teil einer transparenten Kommunikation über ihren proaktiven Datenschutz nutzen.

3.2 Die Regulierung als Innovationskatalysator

Anstatt als bürokratische Hürde zu wirken, treiben die strengen regulatorischen Vorgaben wie DSGVO und AI Act die Entwicklung von sichereren und transparenteren KI-Architekturen wie der lokalen KI voran.[15] Sie zwingen Unternehmen, die Verarbeitung sensibler Daten neu zu überdenken und eine „Compliant-by-Design"-Struktur zu schaffen.[14] Der risikobasierte Ansatz des AI Acts ist hierbei von entscheidender Bedeutung: Er identifiziert Hochrisiko-Anwendungen (z. B. im Gesundheitswesen, bei der Personalverwaltung oder in der Strafverfolgung), deren Scheitern die Gesundheit, Sicherheit oder die Grundrechte von Personen gefährden könnte.[20] In diesen Bereichen ist die lokale Verarbeitung von Daten nicht nur eine Option, sondern ein entscheidender Weg, um die von der Regulierung geforderte Risikominderung von Natur aus zu gewährleisten.[14]

4. Praxisbeispiele und Anwendungsbereiche: Wie Local AI in der Wirtschaft Fuß fasst

Die strategische Relevanz von lokaler KI zeigt sich am deutlichsten in Branchen, die tagtäglich mit besonders sensiblen oder zeitkritischen Daten arbeiten.

4.1 Fallstudien aus datensensitiven Branchen

- Gesundheitswesen: Krankenhäuser nutzen lokale KI-Systeme, um medizinische Bilder zu analysieren oder Diagnosetools zu betreiben.[3] Patientendaten werden direkt vor Ort verarbeitet, sodass sie das Kliniknetzwerk niemals verlassen. Dies gewährleistet die Einhaltung strenger Datenschutzvorschriften wie der DSGVO und der HIPAA in den USA.[14]

- Rechts- und Finanzwesen: Anwaltskanzleien setzen lokale Modelle zur Analyse vertraulicher Dokumente, zur Fallrecherche und zur Vertragsprüfung ein, während Banken Echtzeit-Betrugserkennungssysteme nutzen.[14] Der Schutz von Mandanten- und Kundendaten ist hier oberstes Gebot.[14]

- Produktion und Edge-Computing: In Fertigungsanlagen ermöglichen lokale KI-Systeme Echtzeit-Anwendungen wie die prädiktive Wartung oder die Qualitätskontrolle.[3] Die geringe Latenz ist hier essenziell, da Verzögerungen von nur einer Sekunde Tausende von Dollar an Produktionsfehlern verursachen können.[22]

4.2 Der Aufstieg privater, lokaler Modelle

Die Nachfrage nach datensouveränen Lösungen hat zur Entwicklung von dedizierten, lokalen KI-Plattformen geführt. Ein prominentes Beispiel ist PrivateGPT, eine Software, die es Unternehmen ermöglicht, leistungsstarke generative Sprachmodelle (LLMs) innerhalb ihres eigenen, gesicherten Netzwerks zu betreiben.[23] Solche Lösungen bieten Datenhoheit, Integration mit internen Systemen (wie ERP und CRM) und vollständige Kontrolle über die Daten.[25] Sie sind ein Beweis dafür, dass die Nutzung modernster KI-Technologie nicht mehr zwangsläufig einen Kompromiss beim Datenschutz bedeutet.

4.3 Hybridmodelle: Die Macht des Federated Learning

Eine fortschrittliche Architektur, die die Vorteile von lokaler und Cloud-KI vereint, ist das Federated Learning. Dieser Ansatz ermöglicht das kollaborative Training eines globalen KI-Modells, ohne die Rohdaten der einzelnen Standorte zu zentralisieren.[27] Stattdessen werden nur die Modellparameter, oft verschlüsselt, zwischen den lokalen Datenzentren ausgetauscht. Dadurch bleibt die Privatsphäre der lokalen Datensätze geschützt, während gleichzeitig ein global nutzbares, optimiertes Modell entsteht.[27] Beispiele von Google und NVIDIA zeigen die erfolgreiche Anwendung dieser Methode bei der Entwicklung mobiler KI-Systeme oder bei autonomen Fahrzeugen, wo regionale Datenschutzbestimmungen wie die DSGVO beachtet werden müssen.[27] Dieses Modell löst das Skalierbarkeitsproblem der lokalen KI, indem es die Stärken beider Welten kombiniert und gleichzeitig die Da-

tenhoheit wahrt.

5. Herausforderungen und strategische Überlegungen für die Implementierung

Obwohl die Vorteile der lokalen KI offensichtlich sind, ist ihre Implementierung mit erheblichen Herausforderungen verbunden, die sorgfältiger strategischer Planung bedürfen.

5.1 Hohe Anfangsinvestitionen und Hardware-Anforderungen

Der größte Nachteil der lokalen KI sind die hohen Anfangsinvestitionen in die Hardware. Leistungsstarke GPUs (wie die NVIDIA RTX 4090), ausreichend Arbeitsspeicher (16-64 GB RAM werden empfohlen) und schnelle SSDs sind essenziell für den Betrieb komplexer KI-Modelle.[10] Die Herausforderung beschränkt sich jedoch nicht nur auf den Kauf, sondern auch auf das sogenannte „Right-Sizing" der Hardware.[29] Eine unpassende Konfiguration kann entweder zu Engpässen und Leistungseinbußen führen oder unnötig hohe Kosten verursachen. Dies unterstreicht, dass die Implementierung von Local AI eine genaue Bedarfsanalyse und technische Expertise erfordert.

5.2 Skalierung, Wartung und Betrieb

Die Skalierbarkeit, die Cloud-Lösungen flexibel und dynamisch anbieten, erfordert bei lokalen Systemen physische Upgrades, die kostspielig und zeitaufwändig sind.[3] Darüber hinaus sind Unternehmen bei der lokalen KI für die gesamte Wartung und die Installation von Updates selbst verantwortlich, was bei Cloud-Lösungen vom Anbieter übernommen wird.[5] Dies führt zu einem erhöhten betrieblichen Aufwand und erfordert entweder interne Ressourcen oder die Zusammenarbeit mit erfahrenen Dienstleistern.

5.3 Organisatorische Herausforderungen

Der Übergang zur lokalen KI ist mehr als eine technische Umstellung – er ist ein strategischer Wandel. Unternehmen müssen intern neue Fähigkeiten aufbauen, insbesondere in Bereichen wie Datenwissenschaft, MLOps (Machine Learning Operations) und dem Betrieb von KI-Systemen.[30] Die Integration der KI in bestehende, oft veraltete IT-Systeme kann komplex sein, da Kompatibilitätsprobleme und Kapazitätsengpässe auftreten können.[30] Die erfolgreiche Implementierung erfordert eine enge und funktionsübergreifende Zusammenarbeit zwischen der IT-Abteilung, den Rechtsabteilungen, den Fachbereichen und den Endnutzern.[30] Die Entscheidung für lokale KI ist letztlich eine Entscheidung für Datenhoheit und Unabhängigkeit, aber auch für eine stärkere interne Verantwortung und den strategischen Aufbau von Kernkompetenzen.

6. Fazit & Ausblick: Der Weg zur verantwortungsvollen KI-Strategie

Lokale KI-Systeme sind keine universelle Lösung, sondern eine strategische Alternative, die besonders für datensensitive Anwendungsfälle entscheidende Vorteile in den Bereichen Datenschutz, Kontrolle und Performance bietet. Sie stellen einen praktischen und effektiven Weg dar, die Anforderungen des EU AI Acts an Rechenschaftspflicht und Transparenz proaktiv zu erfüllen und die Kontrolle über die Datenhoheit zurückzugewinnen. Dies ist ein entscheidender Faktor, um das Vertrauen von Kunden und Partnern aufzubauen und langfristig einen Wettbewerbsvorteil zu erzielen.

Die Zukunft der KI-Architekturen wird nicht länger in einer binären Entscheidung zwischen Cloud und lokal liegen. Vielmehr wird der Weg über intelligente, hybride Architekturen führen, die das Beste aus beiden Welten kombinieren.[3] Solche Modelle werden es Unternehmen ermöglichen, sensible Daten sicher lokal zu verarbeiten, während sie für rechenintensive Aufgaben auf die Skalierbarkeit der Cloud zurückgreifen. Die fortschreitende Miniaturisierung von Hardware, wie beispielsweise Neural Processing Units (NPUs), und die Entwicklung kleinerer, effizienterer KI-Modelle (SLMs - Small Language Models) werden die lokale KI in den kommenden Jahren noch zugänglicher und leistungsfähiger machen.[29,5]

Die Wahl der richtigen KI-Strategie ist eine der wichtigsten Entscheidungen für moderne Unternehmen. Sie ist eine Abwägung zwischen Risiko und Innovation, die eine genaue Analyse der eigenen Daten, der regulatorischen Anforderungen und der strategischen Ziele erfordert. Lokale KI ist der Wegbereiter für eine Zukunft, in der KI-Innovation und Vertrauen Hand in Hand gehen.

Martin Wildbacher ist Principal Consultant und Teamlead bei SEQIS.

Er ist in allen Testbereichen tätig: Planung und Koordination, Analyse und Testfallerstellung, Durchführung und Dokumentation sowie Problemmanagement.

Das Arbeiten in einem agilen, dynamischen Umfeld mit wechselnden Teams und Projekten sowie die Freude am Lösen von Herausforderungen sind für ihn die Essenz für den garantierten Erfolg im Softwaretest.

Quellen und weiterführende Informationen:

Hinweis: Dieser Artikel wurde durch Gemini unterstützt.

[1] Consumer Perspectives of Privacy and Artificial Intelligence - IAPP, Zugriff am September 23, 2025, https://iapp.org/resources/article/consumer-perspectives-of-privacy-and-ai/

[2] Cloud AI vs. Local AI: Which Is Best for Your Business? - webAI, Zugriff am September 23, 2025, https://www.webai.com/blog/cloud-ai-vs-local-ai-which-is-best-for-your-business

[3] Local AI vs. cloud AI - which is better for your company? - novalutions, Zugriff am September 23, 2025, https://www.novalutions.de/en/local-ki-vs-cloud-ki-which-suits-your-company-better/

[4] Vergleich von Cloud-basierten KI-Lösungen und On-Premise-Betrieb - Julian Funke, Zugriff am September 23, 2025, https://julian-funke.de/2024/01/24/vergleich-von-cloud-basierten-ki-loesungen-und-on-premise-betrieb/

[5] Choose between cloud-based and local AI models | Microsoft Learn, Zugriff am September 23, 2025, https://learn.microsoft.com/en-us/windows/ai/cloud-ai

[6] Was ist Edge KI? - IBM, Zugriff am September 23, 2025, https://www.ibm.com/de-de/think/topics/edge-ai

[7] Lokale KI Modelle: DSGVO-konforme KI für Unternehmen - docurex, Zugriff am September 23, 2025, https://www.docurex.com/lokale-ki-modelle-2/

[8] AI Architecture Diagram Generator - Eraser IO, Zugriff am September 23, 2025, https://www.eraser.io/ai/architecture-diagram-generator

[9] KI Flussdiagramm erstellen | Miro, Zugriff am September 23, 2025, https://miro.com/de/ai/flussdiagramm-ai/

[10] How to Run AI Locally: Benefits, Challenges, and Solutions - Autonomous, Zugriff am September 23, 2025, https://www.autonomous.ai/ourblog/run-ai-locally

[11] Choosing Between Local AI and Cloud AI: What You Need to Know - Archy.net, Zugriff am September 23, 2025, https://www.archy.net/choosing-between-local-ai-and-cloud-ai-what-you-need-to-know/

[12] Top 10 operational impacts of the EU AI Act – Leveraging GDPR compliance - IAPP, Zugriff am September 23, 2025, https://iapp.org/resources/article/top-impacts-eu-ai-act-leveraging-gdpr-compliance/

[13] Key Issue 5: Transparency Obligations - EU AI Act, Zugriff am September 23, 2025, https://www.euaiact.com/key-issue/5

[14] AI in Data Privacy: How Businesses Can Use Local AI Models to Protect Sensitive Information - AI CERTs, Zugriff am September 23, 2025, https://www.aicerts.ai/news/ai-in-data-privacy-local-models-benefits/

[15] KI und Datenschutz: Wegweiser für einen datenschutzkonformen Einsatz von KI, Zugriff am September 23, 2025, https://www.datenschutzkanzlei.de/ki-und-datenschutz-wegweiser-fuer-einen-datenschutzkonformen-einsatz-von-ki/

[16] Lokale KI-Anwendungen und ihre Rolle im Datenschutz - Mindverse, Zugriff am September 23, 2025, https://www.mind-verse.de/news/lokale-ki-anwendungen-datenschutz

[17] Article 50: Transparency Obligations for Providers and Deployers of Certain AI Systems | EU Artificial Intelligence Act, Zugriff am September 23, 2025, https://artificialintelligenceact.eu/article/50/

[18] How To Use Local AI Models To Improve Data Privacy | The AI Journal, Zugriff am September 23, 2025, https://aijourn.com/how-to-use-local-ai-models-to-improve-data-privacy/

[19] EU Artificial Intelligence Act | Up-to-date developments and analyses of the EU AI Act, Zugriff am September 23, 2025, https://artificialintelligenceact.eu/

[20] AI Act | Shaping Europe's digital future - European Union, Zugriff am September 23, 2025, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[21] Professional Use Cases for Local AI: Privacy-First Solutions in Healthcare, Law, and Business - Enclave AI - Private, Local, Offline AI Assistant for MacOS and iOS, Zugriff am September 23, 2025, https://enclaveai.app/blog/2025/01/16/professional-use-cases-local-ai/

[22] Local AI Agents: A Privacy-First Alternative to Cloud-Based AI, Zugriff am September 23, 2025, https://gloriumtech.com/local-ai-agents-the-privacy-first-alternative-to-cloud-based-ai/

[23] PrivateGPT by Abstracta: Secure GPT Service - Microsoft AppSource, Zugriff am September 23, 2025, https://appsource.microsoft.com/en-us/product/web-apps/abstracta1679321045678.abstracta-privategpt?tab=overview

[24] What Is Private GPT and Why Your Business Might Need It - Litslink, Zugriff am September 23, 2025, https://litslink.com/blog/what-is-private-gpt

[25] Private GPT at Stackscale, Zugriff am September 23, 2025, https://www.stackscale.com/solutions/privategpt/

[26] Fujitsu Private GPT – the GenAI solution with data sovereignty, Zugriff am September 23, 2025, https://www.fujitsu.com/fi/products/data-transformation/private-gpt/

[27] Federated Learning: 5 Use Cases & Real Life Examples, Zugriff am September 23, 2025, https://research.aimultiple.com/federated-learning/

[28] What are the Minimum Hardware Requirements for Artificial Intelligence?, Zugriff am September 23, 2025, https://www.moontechnolabs.com/qanda/minimum-hardware-requirements-for-artificial-intelligence/

[29] Right Sizing Your Edge AI Hardware | OnLogic, Zugriff am September 23, 2025, https://www.onlogic.com/blog/right-sizing-your-edge-ai-hardware/

[30] 5 Challenges of AI Deployment You Can Solve for Your IT Team - NinjaOne, Zugriff am September 23, 2025, https://www.ninjaone.com/blog/challenges-of-ai-deployment/

[31] 6 Common AI Model Training Challenges - Oracle, Zugriff am September 23, 2025, https://www.oracle.com/artificial-intelligence/ai-model-training-challenges/

Die Konvergenz von KI-Ethik und dem AI Act: Von der Abstraktion zur Anwendung

von Alexander Lewisch

Abbildung 1: Quelle: Bild generiert von Gemini (Google AI)

Der EU AI Act kann als bewusster Versuch gesehen werden, klassische philosophisch-ethische Aspekte in verbindliches Recht zu übersetzen, um AI durch Prinzipien wie Autonomie, Verantwortung und Schadensvermeidung zu regulieren. Obwohl dieser Ansatz dem Gesetz eine starke normative Grundlage verleiht, bleibt die Umsetzung durch juristische Lücken und Konflikte mit bestehenden rechtlichen Regelungen unvollständig.

Ethik kodifizieren und Technologie regulieren - eine doppelte Herausforderung

Einleitung - Das Spannungsfeld von Philosophischer Ethik und Recht

Die rasche Entwicklung und Implementierung Künstlicher Intelligenz (KI) aka Artificial Intelligence (AI) hat die Gesellschaft an einen normativen Scheideweg geführt. Grundlegend operieren Ethik und Recht in unterschiedlichen, wenn auch verbundenen, Sphären. Während die Ethik einen Leitfaden für das liefert, was eine Gesellschaft als „wünschenswert" oder „richtig" erachtet, definiert das Recht, was verbindlich und einklagbar ist, gestützt durch Sanktionen. Die jüngste Welle technologischer Innovationen, insbesondere im Bereich der generativen AI, hat die Dringlichkeit einer Konvergenz dieser Sphären dramatisch unterstrichen. Dabei haben nachgewiesene Risiken den Bedarf an strengen Regulierungen und ethischen Praktiken offenkundig gemacht.

Die Europäische Union hat auf diese Herausforderung mit einer Verordnung, dem sogenannten „AI Act", reagiert und damit das weltweit erste umfassende Rechtsinstrument zur Regulierung von AI geschaffen. Dieses Instrument ist jedoch nicht ex nihilo entstanden, sondern es ist das Resultat eines bewussten politischen Weges, der eine zutiefst philosophisch-ethische Komponente beinhaltet. Das erklärte Ziel der Verordnung ist es, die „Einführung einer menschenzentrierten und vertrauenswürdigen künstlichen Intelligenz" zu fördern und gleichzeitig ein „hohes Schutzniveau für Gesundheit, Sicherheit und Grundrechte" zu gewährleisten.[1]

Der italienische Philosoph und Professor für Philosophie und Informationsethik an der University of Oxford Luciano Floridi analysiert diesen regulatorischen Ansatz als inhärent europäisch. In seiner Analyse der Gesetzgebung stellt er fest, dass der AI Act den „grundlegenden Ansatz" der EU-Werte „erbt". Die Vision, die diesem Ansatz zugrunde liegt, ist unverkennbar: Technologie, einschließlich KI, muss im Dienste der Menschheit, ihrer Werte und Bedürfnisse stehen. Floridi kritisiert zwar den Begriff „menschenzentriert" als potenziell „veraltete Terminologie", die anthropozentrische Tendenzen aufweisen könnte, erkennt aber den dahinterliegenden ambitionierten Versuch des AI Acts, eine normative, wertebasierte Vision in bindendes, operationales Recht zu gießen.[2]

Somit kann der AI Act als juristisches Endprodukt eines nicht nur mehrjährigen, organisierten politischen, sondern auch philosophischen Prozesses verstanden werden. Dieser Prozess begann formell mit der Einsetzung einer unabhängigen hochrangigen Expertengruppe für KI, der "High-Level Expert Group on AI", kurz HLEG, durch die Europäische Kommission im Jahr 2018.[3] Die HLEG legte 2019 ihre Ethik-Leitlinien vor, die als direkte Blaupause für den Gesetzesvorschlag der Kommission von 2021 und die finale Verordnung von 2024 dienten. Diese Pfadabhängigkeit von der Ethik zum Recht ist zugleich die größte Stärke und die größte Schwäche des AI Acts. Sie ist eine Stärke, da sie dem Gesetz eine kohärente normative Grundlage verleiht, die in den europäischen Grundwerten verankert ist.

Sie ist jedoch auch eine potenzielle Schwäche, da der AI Act nun daran gemessen wird, ob seine juristischen Mechanismen, wie bspw. in Artikel 14 "zur menschlichen Aufsicht", die hohen ethischen Ziele tatsächlich erreichen können, oder ob sie im regulatorischen Vollzug zur bürokratischen Bedeutungslosigkeit verkommen.

Philosophische Ethik und ihre Relevanz für AI

Die Debatte über AI-Ethik kann im Kern als eine Anwendung klassischer philosophischer Theorien auf neue technologische Gegebenheiten verstanden werden. Der AI Act und die ihm vorangegangenen Ethik-Leitlinien sind tief in den klassischen Theorien philosophischer Traditionen (Tugendethik, Deontologie und Utilitarismus) verwurzelt.

Tugendethik und Verantwortung

Die aristotelische Tugendethik argumentiert, dass Ethik eine Frage des Charakters und der praktischen Weisheit des Handelnden ist. Eine dementsprechende AI-Entwicklung erfordert strukturierte ethische Überlegungen. Der AI Act erzwingt solche tugendethischen Prozesse, indem er beispielsweise Risikomanagement und Qualitätsmanagement vorschreibt, die über den gesamten Lebenszyklus des KI-Systems aufrechterhalten werden müssen. Hierzu lieferte die bekannte Publizistin Hannah Arendt in ihrem Werk "Eichmann in Jerusalem: Ein Bericht von der Banalität des Bösen" (1964)[4] eine eindringliche Warnung, warum eine institutionalisierte Verantwortungsethik notwendig ist. Ihre Analyse der „Banalität des Bösen" zeigt die Verantwortungslosigkeit, die entsteht, wenn Individuen ihre moralische Urteilskraft an bürokratische Prozesse delegieren und ihre persönliche Verantwortung nicht wahrnehmen. Arendt sieht in der individuellen Verantwortung das zentrale Kriterium für die Bewahrung der Urteilsfähigkeit.

Die größte „banale" Gefahr bei der Anwendung von KI ist der „Automation Bias", die unkritische Übernahme von algorithmischen Empfehlungen. Hierfür liefert Arendt die philosophische Begründung, warum die „Menschliche Aufsicht" gemäß Artikel 14 des AI Acts so entscheidend ist. Artikel 14 ist der juristische Abwehrmechanismus gegen die „Banalität des Algorithmus", indem er dem Menschen explizit die Fähigkeit und die Pflicht zuweist, die Ausgabe des Systems zu ignorieren, abzuändern oder umzukehren, und somit zum Instrument der Rückforderung der persönlichen Verantwortung fungiert.

Deontologie und Autonomie

Die europäische Grundrechtstradition, die im AI Act eine zentrale Rolle einnimmt, ist ohne die deontologische Ethik Immanuel Kants kaum denkbar. Im Zentrum der kantischen Deontologie steht nicht der Nutzen einer Handlung, sondern die Pflicht, die sich aus der Vernunft ergibt. Das höchste Prinzip dieser Ethik ist die Autonomie des Willens. Kant nennt dies „das alleinige Prinzip der Moral" (Kant, GMS, S. 67).[5] Aus dieser Autonomie leitet Kant die „Selbstzweckformel" des Kategorischen Imperativs (Kant, GMS, S. 36ff) ab, die gebietet, die Menschheit, sowohl in der eigenen Person als auch in der Person eines jeden anderen, „jederzeit zugleich als Zweck, niemals bloß als Mittel" zu gebrauchen.[6] Die HLEG-Leitlinien spiegeln diesen Gedanken direkt wider, wenn sie betonen, dass der „Respekt vor der Menschenwürde" erfordert, dass Menschen „niemals bloß als Objekte (...) sortiert, bewertet, (...) oder manipuliert werden".[7]

Die Verbote in Artikel 5 des AI Acts sind ein direktes juristisches Abbild dieser kantischen Ethik. Das Verbot von KI-Systemen, die subliminale oder manipulative Techniken einsetzen, oder das Verbot von „Social Scoring", ist nicht primär utilitaristisch begründet. Diese Systeme werden nicht verboten, weil sie mehr schaden als nutzen, sondern weil sie, im Sinne Kants, per se die Autonomie und menschliche Würde verletzen und Menschen bloß als Mittel zum Zweck sehen, d.h. zur Erreichung eines externen Ziels, wie etwa die Steuerung des Verhaltens oder soziale Konformität.

Konsequentialismus und Schadensvermeidung

Während die kantische Deontologie die absoluten Verbote des AI Acts erklärt, liefert der Konsequentialismus, insbesondere der Utilitarismus von Jeremy Bentham oder John Stuart Mill, die philosophische Blaupause für den regulativen Teil des Gesetzes, d.h. den risikobasierten Ansatz. Der Konsequentialismus bewertet die Moral einer Handlung anhand ihrer Folgen. John Stuart Mills „Harm Principle" (Schädigungsprinzip), dargelegt in seinem Werk "Über die Freiheit"[8], besagt: „Dieses Prinzip lautet, dass der einzige Zweck, der die Menschheit berechtigt, vereinzelt oder vereinigt, jemandes Handlungsfreiheit zu beeinträchtigen, der Selbstschutz ist; dass der einzige Zweck, der rechtfertigt, Macht über irgendein Mitglied einer zivilisierten Gemeinschaft gegen seinen Willen auszuüben, der ist, die Schädigung anderer zu verhüten. Sein eigenes Wohl, das leibliche wie das moralische, ist kein ausreichender Grund dafür" (Mill, 1859, S. 316).

Das besagte Schädigungsprinzip ist durch den vierstufigen, risikobasierten Ansatz des AI Acts präzise operationalisiert:

- Mills Libertarismus (Keine Intervention): Der letzte Teil des oben genannten Zitats zeigt, dass das eigene Wohl also kein ausreichender Grund für eine Beschränkung ist. Dies entspricht den Kategorien für KI-Systeme, die laut AI Act nur ein minimales oder gar kein Risiko aufweisen und weitgehend unreguliert sind. Hier überwiegt die Freiheit der Innovation.
- Mills Schadensprinzip (Intervention): Das Zitat beinhaltet auch das

Recht einer Gesellschaft, eines Staates oder wie es Mill nennt, einer zivilisierten Gemeinschaft, einzugreifen, um die Schädigung anderer abzuwenden. Dies entspricht den regulierten Kategorien des AI Acts. Sobald ein KI-System „ernsthafte Risiken für die Gesundheit, Sicherheit oder Grundrechte" Dritter darstellt, wird es als „Hochrisiko-System" (High-Risk) eingestuft und strengen Verpflichtungen unterworfen. Der AI Act ist in diesem Sinne eine Verordnung, die sicherstellt, dass KI Sicherheit, Gesundheit und Grundrechte wahrt.

Die Entwicklung europäischer KI-Ethik-Leitlinien

Der AI Act ist, wie in der Einleitung dargelegt, das Endprodukt eines Prozesses, der mit ethischen Überlegungen begann. Die 2018 von der Europäischen Kommission eingesetzte High-Level Expert Group on AI (HLEG) hatte den Auftrag, einen ethischen Rahmen für die KI-Entwicklung in Europa zu definieren. Im April 2019 präsentierte die HLEG ihre Ethik-Leitlinien für eine vertrauenswürdige KI. Diese definieren vertrauenswürdige AI als ein System, das drei Komponenten aufweisen muss, die idealerweise in Harmonie zusammenwirken. Die 3 Säulen der HLEG sind[9]:

1. Rechtmäßig (Lawful): Achtung aller geltenden Gesetze und Vorschriften.
2. Ethisch (Ethical): Achtung ethischer Prinzipien und Werte.
3. Robust (Robust): Sowohl aus technischer als auch aus sozialer Perspektive.

Der AI Act ist die Kodifizierung der ersten Säule (rechtmäßig), die sich jedoch inhaltlich massiv auf die zweite Säule (ethisch) stützt, um die dritte (robust) zu gewährleisten.

Als normatives Fundament identifizierte die HLEG zusätzlich vier ethische Grundsätze[10], die auf den diskutierten philosophischen Traditionen und den EU-Grundrechten basieren:

1. Achtung der menschlichen Autonomie
2. Schadensvermeidung
3. Fairness
4. Erklärbarkeit

Der entscheidende Schritt der HLEG war die Übersetzung dieser vier abstrakten Prinzipien in sieben konkrete, operative Schlüsselanforderungen. Diese sieben HLEG-Anforderungen bilden die Blaupause für das Herzstück des AI Acts, d.h. die Verpflichtungen für Hochrisiko-Systeme in Kapitel III. Sie wurden in der „Assessment List for Trustworthy AI (ALTAI)" operationalisiert, einer praktischen Checkliste für Entwickler und Anwender[11]. Die Konvergenz von Ethik und Recht wird am deutlichsten, wenn man die sieben HLEG-Anforderungen direkt mit den spezifischen Artikeln des AI Acts für Hochrisiko-Systeme vergleicht (in Klammer). Zu den HLEG-Anforderungen zählen:

1. Menschliche Handlungsfähigkeit und Aufsicht
   (**Artikel 14**: Menschliche Aufsicht)
2. Technische Robustheit und Sicherheit
   (**Artikel 15**: Genauigkeit, Robustheit und Cybersicherheit)
3. Privatsphäre und Daten-Governance
   (**Artikel 10**: Daten und Data Governance)
4. Transparenz
   (**Artikel 13**: Transparenz und Bereitstellung von Informationen)
5. Vielfalt, Nichtdiskriminierung und Fairness
   (**Artikel 10**: Daten und Data Governance)
6. Gesellschaftliches Wohlergehen und ökologische Nachhaltigkeit
   (**Artikel 95**: Verhaltenskodizes für die freiwillige Anwendung von spezifischen Anforderungen)
7. Rechenschaftspflicht
   (**Artikel 11**: Technische Dokumentation und **Artikel 12**: Protokollierung).

Wo das Recht hinter der Ethik zurückbleibt

Der AI Act ist keine perfekte Übersetzung, da in der Transformation von „weicher" Ethik zu „hartem" Recht zwangsläufig Spannungen, Lücken und Zielkonflikte entstehen.

Spannungsfeld 1: AI Act vs. DSGVO (Datenschutz-Grundverordnung)

Ein primäres Spannungsfeld und ein fundamentaler Zielkonflikt besteht im Verhältnis zur Datenschutz-Grundverordnung (DSGVO). Das ethische Ziel der „Fairness" verlangt nach der rechtlichen Pflicht aus Art. 10 AI Act, Voreingenommenheit (Bias) zu erkennen und zu korrigieren. Um jedoch Bias (z. B. aufgrund des Geschlechts oder der ethnischen Herkunft) effektiv zu erkennen, müssen Modelle oft mit sensiblen Daten (gem. Art. 9 DSGVO), die eben diese Merkmale beschreiben, trainiert oder getestet werden.[12]

Abbildung 2: Quelle: Bild generiert von Gemini (Google AI)

Hier kollidiert das ethische Ziel der Fairness (AI Act) frontal mit dem ethischen Ziel des Datenschutzes (DSGVO), das die Verarbeitung solcher Daten grundsätzlich verbietet. Der AI Act versucht diesen Konflikt in Art. 10 Abs. 5 aufzulösen. Er schafft eine streng zweckgebundene Ausnahme, die die Verarbeitung sensibler Daten nur zum Zweck der Bias-Erkennung und Bias-Korrektur erlaubt, und nur, wenn dies unbedingt erforderlich ist und durch strenge technische und organisatorische Maßnahmen (z. B. Pseudonymisierung, Zugriffskontrollen) geschützt wird. Diese Lösung offenbart jedoch eine kritische Lücke, da diese Ausnahme in Art. 10 Abs. 5 ausschließlich für Anbieter von Hochrisiko-KI-Systemen gilt. Sie gilt nicht für Anbieter von General-Purpose AI (GPAI) Modellen oder Systemen mit geringerem Risiko, obwohl auch diese ein Diskriminierungspotenzial aufweisen. Hier bleibt die rechtliche Lösung (AI Act) hinter der ethischen Forderung (Fairness in allen Systemen) signifikant zurück.

Spannungsfeld 2: AI Act vs. DSA (Digital Services Act)

Ein weiterer Konflikt entsteht im Zusammenspiel mit dem Digital Services Act (DSA), insbesondere bei der Regulierung von KI-Systemen, die von sehr großen Online-Plattformen eingesetzt werden. Das ethische Prinzip der Rechenschaftspflicht (Accountability) erfordert Auditierbarkeit. Um KI-Systeme auf systemische Risiken zu prüfen, benötigen unabhängige Forscher und die Zivilgesellschaft Zugang zu relevanten Daten. Der DSA erkennt diese Notwendigkeit an und gewährt in Artikel 40 geprüften Forschern explizit einen solchen Datenzugang für die Analyse systemischer Risiken. Der AI Act hingegen, der ebenfalls systemische Risiken, insbesondere durch GPAI, adressiert, enthält keine vergleichbare Bestimmung.[13] Dies stellt einen schwerwiegenden Mangel in der Kohärenz der EU-Digitalgesetzgebung dar. Der AI Act behindert hier potenziell die Erreichung seiner eigenen ethischen Ziele, indem er Forschern den Zugang verwehrt, den ein anderes EU-Gesetz (der DSA) bereits als fundamental für die Aufsicht erachtet hat.

Spannungsfeld 3: Operationalisierung von „Accountability"

„Accountability" (Rechenschaftspflicht) ist ein zentrales ethisches Prinzip. Im AI Act wird es primär durch prozedurale Pflichten durch die Artikel 11 (Technische Dokumentation) und Artikel 12 (Protokollierung) operationalisiert. Anbieter müssen ex ante dokumentieren, wie ihr System konzipiert ist, und ex post sicherstellen, dass sein Betrieb durch „Logs" nachvollziehbar ist. Kritiker argumentieren jedoch, dass dies „Accountability" auf eine reine Verfahrensdokumentation reduziert.[14]

Diese entscheidende Haftungslücke (Liability) sollte durch die „AI Liability Directive" (KI-Haftungsrichtlinie) geschlossen werden. Dieser Richtlinienentwurf, der darauf abzielte, die Beweislast für Opfer von KI-Schäden zu erleichtern, wurde jedoch von der Europäischen Kommission im Februar 2025 zurückgezogen. Dies hinterlässt die größte Lücke im System. Ohne ein klares Haftungsregime bleibt die Rechenschaftspflicht ein ethisches Postulat ohne die notwendige juristische Macht, um im Schadensfall Gerechtigkeit zu gewährleisten.[15]

Fazit: Eine verantwortungsvolle Zukunft durch die Symbiose von Ethik und Recht

Der AI Act ist nicht bloß ein weiterer Rechtsakt der Digitalregulierung, sondern er ist durch den ethischen Anspruch auch eine kodifizierte Philosophie, weil er versucht, die normativen Grundlagen der europäischen Ethik in verbindliche, technische und prozessuale Verpflichtungen für Hochrisiko-Systeme zu übersetzen. Ethik und Recht gehen im AI Act insofern Hand in Hand, als die ethischen Leitlinien der HLEG die direkte Vorlage für die juristischen Kernanforderungen in den Artikeln 10-15 lieferten. Während die Ethik für das „Warum" (z. B. Schutz der Autonomie) sorgt, steuert das Recht „Wie" (z. B. die menschliche Aufsicht in Artikel 14) bei. Bei aller Zuversicht ist die Symbiose jedoch unvollständig und von Friktionen geprägt. Es verbleiben signifikante Spannungen in der Kohärenz mit bestehendem Recht (DSGVO, DSA). Die schwerwiegendste Lücke klafft jedoch im Bereich der Haftung, da durch das Zurückziehen der KI-Haftungsrichtlinie die Operationalisierung von „Accountability" auf prozedurale Dokumentation beschränkt bleibt.

Quellen und weiterführende Informationen:

[1] European Commission: Excellence and trust in artificial intelligence. In: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/excellence-and-trust-artificial-intelligence_en

[2] Floridi, Luciano (2021): The European Legislation on AI: A Brief Analysis of its Philosophical Approach. SSRN Electronic Journal. 10.2139/ssrn.3873273.

[3,7] High-Level Expert Group on Artificial Intelligence (2018): Ethics Guidelines for trustworthy AI. In: https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf

[4] Arendt, Hannah: Eichmann in Jerusalem. Ein Bericht von d. Banalität d. Bösen, Piper, München 1964.

5,6] Kant, Immanuel: Grundlegung zur Metaphysik der Sitten (Vol. 28). L. Heimann, Berlin 1870.

[8] Mill, John Stuart (1859): Über die Freiheit. In: Schefzcyk, Michael & Schmidt-Petri, Christoph (Hg.): John Stuart Mill: Ausgewählte Werke, Band III/1 Freiheit, Fortschritt und die Aufgaben des Staates. Individuum, Moral und Gesellschaft, Wachholtz Verlag, Kiel/Hamburg 2021.

[9,11] European Commission: Ethics guidelines for trustworthy AI. In: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[10] Peets, Lisa; Hansen, Marty & Choi, Sam Jungyun: EU High-Level Working Group Publishes Ethics Guidelines for Trustworthy AI. In: https://www.covingtondigitalhealth.com/2019/04/eu-high-level-working-group-publishes-ethics-guidelines-for-trustworthy-ai/

[12,13] Hacker, Philipp: Der AI Act im Spannungsfeld von digitaler und sektoraler Regulierung. Hrsg. Bertelsmann Stiftung. Gütersloh 2024. In: https://www.bertelsmann-stiftung.de/fileadmin/files/user_upload/AI_Act_im_Spannungsfeld__2024_final.pdf

[14] Oduro, Serena, Moss, Emanuel & Metcalf, Jacob (2022). Obligations to assess: Recent trends in AI accountability regulations. Patterns (New York, N.Y.), 3(11), 100608. https://doi.org/10.1016/j.patter.2022.100608

[15] Gaudszun, Timo; Shin, Jeffrey & Parsons, Natasha: AI Watch: Global regulatory tracker - European Union. In: https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-european-union

Alexander Lewisch ist Consultant bei SEQIS.

Als IT-Quereinsteiger und Wiedereinsteiger konnte er Erfahrungen im Software Testing sammeln und Einblicke in diverse Bereiche wie Testmanagement, Testdurchführung, Testautomation, Testkoordination, Reporting und Consulting sammeln.

Derzeit widmet er sich verschiedenen Testprojekten im Bereich der Remote Testing Services (RTS).

KI im Unternehmen: Ein Praxis-Leitfaden zu Datenschutz, AI Act und Compliance

von Emre Kurtulush

Abbildung 1: Quelle: Bild generiert von Gemini (Google AI)

Im Jahr 2025 ist Künstliche Intelligenz (KI) ein fundamentaler Pfeiler der Unternehmensstrategie. Doch nach dem ersten Hype stehen Unternehmen im deutschsprachigen Raum (DACH-Region) vor kritischen, strategischen Fragen, die über die reine Leistung der Modelle hinausgehen.

Für welchen Bereich sind diese Modelle sicher einzusetzen – für das Development, die Texterzeugung oder die Testfallerstellung? Was sind die tatsächlichen Vor- & Nachteile der verschiedenen KI-Modelle im operativen Einsatz?

Vor allem aber: Was passiert mit Ihren sensiblen Unternehmensdaten? Werden Ihre Daten zur weiteren Entwicklung des KI-Modells verwendet? Wie sicher sind Ihre Daten wirklich? Wo werden diese gespeichert – in der EU oder potenziell in Drittstaaten? Und wer genau kann Ihre Prompts sehen?

Die Antworten auf diese Fragen definieren die „rote Linie" zwischen einer konformen, nachhaltigen KI-Implementierung und einem untragbaren rechtlichen Risiko. In diesem Artikel beantworten wir genau diese Fragen.

1 Eine Taxonomie moderner KI-Modelle

Um fundierte Entscheidungen treffen zu können, ist eine klare Abgrenzung der verfügbaren Modellkategorien erforderlich.

1.1 Große Sprachmodelle (LLMs): Das Fundament

Große Sprachmodelle (Large Language Models, LLMs) bilden die Grundlage der aktuellen generativen KI-Revolution. Sie werden auf Basis riesiger Text- und Datensätze trainiert und nutzen Deep-Learning-Techniken, um menschliche Sprache zu verarbeiten und zu generieren. Ihre Kernkompetenzen umfassen eine Vielzahl von Aufgaben der natürlichen Sprachverarbeitung (NLP), darunter Textzusammenfassung, Maschinenübersetzung, Beantwortung von Fragen und das Verfassen kreativer Texte.

Zu den wichtigsten Akteuren in diesem Bereich gehören die GPT-Modellfamilie (Generative Pre-trained Transformer) von OpenAI[6], Gemini von Google, die Claude-Modellfamilie von Anthropic, die sich auf Sicherheit und komplexe Schlussfolgerungen konzentriert, und die Llama-Modellfamilie von Meta. [5]

1.2 Spezialisierte generative Modelle

Neben den generalistischen LLMs hat sich ein Markt für hochspezialisierte Modelle entwickelt, die für bestimmte Aufgaben optimiert sind.

Code-Generierung und -Analyse

Diese Modelle sind speziell auf Programmiersprachen, Software-Bibliotheken und Entwicklungskonzepte trainiert.[8] Sie dienen als „Pair-Programmer" und verstehen den spezifischen Kontext einer Codebasis, einschließlich Dateistrukturen, Importabhängigkeiten und Namenskonventionen. [10]

Das prominenteste Beispiel ist GitHub Copilot, das eine Suite verschiedener Modelle nutzt, darunter GPT-4.1 und Claude 3.5 Sonnet.[4] Andere spezialisierte Code-Modelle umfassen Devstral[2] und Grok Code Fast.[11] Das deutsche Bundesamt für Sicherheit in der Informationstechnik (BSI) sieht in der „Analyse und Härtung von Programmcode" eine wesentliche Chance dieser Technologie.[8]

Bild-, Video- und Audiogeneratoren

Diese Modelle operieren nicht primär mit Text, sondern sind auf die Erzeugung und Bearbeitung visueller oder auditiver Daten spezialisiert.[8] Bekannte Beispiele sind Bildgeneratoren wie DALL-E 3 und Midjourney, Video-Generatoren wie OpenAI's Sora [15] und Pika Labs sowie Stimm-Generatoren wie ElevenLabs. Ihre Anwendungsbereiche reichen von Marketing und Design bis hin zur Erstellung synthetischer Trainingsdaten.[2]

1.3 Multimodale Modelle: Die nächste Evolutionsstufe

Die jüngste Generation von KI-Modellen ist multimodal, was bedeutet, dass sie mehrere Arten von Informationen (Modalitäten) gleichzeitig verarbeiten können, wie z.B. Text, Bilder, Audio und Code.[1]

Modelle wie GPT-4o[3], Google Gemini [4] und Alibaba Qwen 2.5 Omni[2] können komplexe Anfragen bearbeiten, die verschiedene Datentypen kombinieren. Sie können beispielsweise ein Diagramm (Bild) analysieren, den darin enthaltenen Code (Text) extrahieren und eine mündliche Zusammenfassung (Audio) liefern. Diese Fähigkeit eröffnet neue Anwendungsfälle in der Datenanalyse, Barrierefreiheit und im kreativen Design.

1.4 Open-Source vs. proprietäre Modelle: Ein fundamentaler Unterschied

Die strategisch wichtigste Unterscheidung für Unternehmen ist die zwischen proprietären und Open-Source-Modellen.

- Proprietäre Modelle: Dies sind geschlossene („Closed Source") Modelle, die von Unternehmen wie OpenAI, Anthropic und Google entwickelt und als kommerzieller Dienst (typischerweise über eine API) angeboten werden. Die Modellarchitektur und die Trainingsdaten sind Geschäftsgeheimnisse.
- Open-Source-Modelle: Bei diesen Modellen sind die Modellgewichte und oft auch der Trainingscode öffentlich verfügbar.[2] Prominente Beispiele sind die Llama-Modellfamilie von Meta und die Modelle von Mistral AI.[1]

Diese Wahl stellt eine fundamentale Weichenstellung für die Daten-Governance:

- Pfad A (Proprietär): Unternehmen, die eine proprietäre API nutzen, lagern das technische Betriebs- und Datenschutzrisiko an den Anbieter aus. Sie sind auf dessen vertragliche Zusicherungen (z.B. im Auftragsverarbeitungsvertrag), dessen Sicherheitsarchitektur und dessen Umgang mit Datentransfers in Drittstaaten angewiesen. Dies erfordert eine tiefgehende juristische Prüfung der Anbieterrichtlinien (siehe Kapitel 4).
- Pfad B (Open-Source): Unternehmen, die ein Open-Source-Modell auf ihrer eigenen Infrastruktur betreiben (On-Premise oder in einer Private Cloud in der EU)[17], übernehmen das Risiko und die Kontrolle vollständig selbst. Der Vorteil ist die maximale Datenhoheit: Sensible Daten verlassen zu keinem Zeitpunkt die kontrollierte Umgebung des Unternehmens.[19] Die Fragen nach Datentraining, Speicherort und Einsichtnahme durch Dritte werden trivial beantwortet („findet nicht statt" oder „nur intern"). Der Nachteil sind die signifikant höheren Kosten und die technische Komplexität für Implementierung, Wartung und Feinabstimmung.[2]

2 Anwendungsbereiche und praktische Einsatzszenarien

Die kategorisierten Modelle ermöglichen eine breite Palette von Anwendungsfällen, deren Eignung von der jeweiligen Datensensitivität abhängt.

2.1 Softwareentwicklung und IT-Betrieb

Der IT-Sektor profitiert massiv von spezialisierten KI-Modellen.

- Automatisierte Code-Erstellung: KI-gestützte „Pair-Programmer" wie GitHub Copilot[9] beschleunigen die Softwareentwicklung erheblich, indem sie Code-Blöcke in Echtzeit vorschlagen und Routineaufgaben automatisieren.[20]
- Debugging und Code-Analyse: Modelle mit tiefem logischem Verständnis (z.B. Claude 3.7 Sonnet, GPT 4.5) werden zur Fehlersuche in komplexen Systemen eingesetzt.[4] Das BSI hebt die Fähigkeit zur „Analyse und Härtung von Programmcode" als strategische Chance hervor.[8]
- Generierung von Testfällen: Ein hoch-rentabler Anwendungsfall ist die automatisierte Erstellung von Software-Tests. Basierend auf realen Problemstellungen wurde ein 3-Schritte-Prozess entwickelt, der die Effizienz steigert[22]:

1. Input (Prosa): Eine Anforderung in natürlicher Sprache (z.B. eine User Story) wird dem LLM übergeben.
2. Output (BDD): Das Modell generiert daraus strukturierte Akzeptanzkriterien im BDD-Format (Behaviour-Driven Development), z.B. „Given-When-Then".
3. Output (Code): Das LLM transformiert die BDD-Fälle in ein lauffähiges Test-File (z.B. Cucumber) und generiert den fertigen Automatisierungscode (z.B. in JavaScript).

2.2 Texterzeugung und Unternehmenskommunikation

Dies ist der klassische Anwendungsbereich von LLMs, der von internen Prozessen bis hin zum externen Marketing reicht.

- Marketing und Content-Erstellung: KI-Modelle generieren Marketingtexte, Blog-Beiträge, Social-Media-Updates und Werbeinhalte. Spezialisierte Tools wie Jasper oder Copy.ai nutzen diese Kernmodelle. Gekoppelt mit Bildgeneratoren wie DALL-E oder Midjourney[14] können ganze Kampagnen erstellt werden.[2]
- Interne Dokumentation und Wissensmanagement: Die Effizienz im Büroalltag wird durch die automatisierte Zusammenfassung von Besprechungsnotizen [10], den Entwurf von E-Mails[2] und die Erstellung technischer Dokumentationen[26] gesteigert.
- Automatisierter Kundenservice: LLMs sind die treibende Kraft hinter fortschrittlichen Chatbots und Unterhaltungsagenten, die Kundenanfragen in natürlicher Sprache beantworten können.[2]

2.3 Strategische Analyse und Forschung

Modelle mit großen Kontextfenstern und Echtzeit-Internetzugang ermöglichen neue Formen der strategischen Analyse.

- Domänenspezifische Dokumentenanalyse: Modelle wie Claude 2, die bis zu 100.000 Tokens (entspricht 100+ von Seiten) verarbeiten können, werden zur Analyse umfangreicher Fachdokumente eingesetzt.[26] Anwendungsfälle umfassen die Prüfung juristischer Verträge oder die Interpretation medizinischer Studiendaten.[2]
- Markt- und Trendforschung: Modelle mit integriertem Echtzeit-Webzugriff (z.B. Perplexity, Grok oder das neue Claude-Feature) können genutzt werden,

um Live-Daten zu überwachen, Markt- und Wettbewerbstrends zu analysieren und Recherchen mit direkten Quellenangaben durchzuführen.[2]

Die Vielfalt dieser Anwendungsfälle, von der Erstellung öffentlicher Marketingtexte bis zur Analyse streng vertraulicher Rechtsdokumente[2] oder dem Debugging von Code, der personenbezogene Daten verarbeitet [4], macht deutlich, dass ein Unternehmen nicht ein KI-Modell wählen kann. Es muss eine gestaffelte KI-Strategie entwickelt werden, die auf der Datensensitivität des jeweiligen Anwendungsfalls basiert. Eine mögliche Architektur könnte wie folgt aussehen:

- Stufe 1 (Öffentliche Daten): Einsatz von Consumer-Tools für nicht-vertrauliche Aufgaben (z.B. allgemeine Recherche, Brainstorming für öffentliche Blogposts).
- Stufe 2 (Vertrauliche/Proprietäre Daten): Zwingender Einsatz von Enterprise/API-Versionen mit Auftragsverarbeitungsvertrag (AVV) für interne Dokumente, E-Mails oder die Entwicklung von nicht-kritischem Code. [10]
- Stufe 3 (Streng geheime/Personenbezogene Daten (PII/DS-GVO)): Zwingender Einsatz von Diensten mit garantierter EU-Datenresidenz (z.B. Google Vertex AI in einer EU-Region[29]) oder selbst-gehosteten Open-Source-Modellen.[17]

Die Beantwortung von Fragen nach Anwendungsbereiche definiert somit direkt die Compliance-Anforderungen für die Beantwortung der Fragen über Datenschutz.

3 Chancen und Risiken im Überblick

Der Einsatz von KI-Modellen birgt ein duales Potenzial von erheblichen Vorteilen und ebenso signifikanten Risiken.

3.1 Vorteile: Produktivität, Innovation und Skalierbarkeit

Die primären Vorteile des KI-Einsatzes sind operativer und strategischer Natur.

- Produktivitätssteigerung und Zeitersparnis: Die Automatisierung von Routineaufgaben, sei es in der Softwareentwicklung[20] oder bei der Testerstellung[22], ist der größte Hebel. Entwickler können sich auf komplexe Probleme konzentrieren, anstatt Standardcode zu schreiben. Studien zeigen eine Reduzierung der Testerstellungszeit um 60-80%.[31]
- Qualitätsverbesserung und Fehlervermeidung: Im Software-Engineering helfen LLMs, die Fehlerquote im Code zu senken[20] und eine konsistente Code-Qualität sicherzustellen.[22] KI-gestützte Testtools sind in der Lage, Randfälle und Sicherheitslücken zu identifizieren, die menschliche Tester oft übersehen.[31]
- Wissensdemokratisierung und Skalierbarkeit: Durch die Kombination von Low-Code-Plattformen mit generativer KI können auch Mitarbeiter ohne tiefgehende Programmierkenntnisse („Citizen Developer“) funktionale Anwendungen erstellen. Dies entlastet die zentralen IT-Abteilungen und beschleunigt die Digitalisierung im gesamten Unternehmen.[32]

3.2 Nachteile und inhärente Risiken

Diesen Vorteilen stehen technische, operative und juristische Risiken gegenüber.

- Technische Limitierungen: Die Modelle sind nicht fehlerfrei.
- Halluzinationen: Ein bekanntes Problem ist das „Halluzinieren“ – das Generieren von plausibel klingenden, aber sachlich falschen oder unsinnigen Informationen.[12] Dies stellt ein kritisches Risiko dar, wenn die Modelle für Analyse- oder Entscheidungsfindungsaufgaben eingesetzt werden.
- Latenz und Kosten: Die leistungsfähigsten Modelle (z.B. GPT-4.5) sind oft langsamer (höhere Latenz) und teurer im Betrieb, was ihren Einsatz in Echtzeitanwendungen einschränken kann.[24]
- Sicherheitsrisiken (Operativ und Implementierung): Das BSI warnt vor spezifischen Gefahren während des gesamten Lebenszyklus der Modelle.[8]
- Prompt Injection: Angreifer können die Modelle durch manipulierte Eingaben (Prompts) dazu bringen, ihre internen Schutzmaßnahmen zu umgehen, bösartige Inhalte zu generieren oder sensible Informationen preiszugeben.
- Risiko der Code-Analyse: Einer der größten Vorteile – die Fähigkeit, Code auf Fehler zu analysieren[8] – birgt gleichzeitig eines der größten Risiken.

Dieser Widerspruch, das „Analyse-Paradoxon“, stellt ein neues, kritisches Einfallstor für Daten-Exfiltration dar. Die Kausalkette dieses Risikos ist wie folgt:

1. Ein Entwickler möchte eine proprietäre Anwendung auf Schwachstellen prüfen und übergibt den vollständigen Quellcode an einen KI-Dienst (z.B. die OpenAI-API) mit dem Prompt: „Analysiere diesen Code auf Sicherheitslücken.“
2. In diesem Moment wird das wertvollste geistige Eigentum (IP) des Unternehmens zusammen mit einer potenziellen Liste seiner Schwachstellen an einen Drittanbieter (z.B. OpenAI oder Anthropic) übertragen.
3. Die Datenübertragung erfolgt in die USA, wo die Speicherung stattfindet.[33]
4. Selbst, wenn die Enterprise-API-Richtlinie ein Training mit diesen Daten verbietet (was der Fall ist), werden die Daten zur Missbrauchsüberwachung (Abuse Monitoring) standardmäßig für 30 Tage gespeichert.[35]
5. Während dieser 30 Tage sind die

Daten – der Quellcode – potenziell der Einsichtnahme durch Mitarbeiter des Anbieters und dem Zugriff durch US-Behörden (im Rahmen des CLOUD Act) ausgesetzt.[37]

Dieser Vorgang macht aus einem gut gemeinten Sicherheits-Audit einen potenziell katastrophalen IP-Diebstahl oder Sicherheitsvorfall. Das „Analyse-Paradoxon" zeigt zwingend, dass sicherheitskritische Anwendungsfälle (wie die Analyse von proprietärem Code) ausschließlich auf einer Stufe-3-Architektur (On-Premise Open-Source)[17] stattfinden dürfen.

4 Kritische Analyse der Daten-Governance

Hier beantworten wir die kritischen Datenschutzfragen des Nutzers, analysieren die Richtlinien der Hauptanbieter (OpenAI, Anthropic, Google, Meta) und ziehen die entscheidende Trennlinie zwischen Consumer- und Enterprise-Diensten.

4.1 Die rote Linie: Warum Consumer-Tools und Enterprise-APIs rechtlich völlig unterschiedlich zu bewerten sind

Sobald ein Unternehmen im DACH-Raum eine KI zur Verarbeitung von Daten einsetzt, die sich auf Mitarbeiter oder Kunden beziehen (personenbezogene Daten), unterliegt es der DSGVO. In diesem Szenario agiert das Unternehmen als „Verantwortlicher" (Controller) und der KI-Anbieter (z.B. OpenAI) als „Auftragsverarbeiter" (Processor).

Nach Artikel[28] der DSGVO ist für diese Konstellation der Abschluss eines Auftragsverarbeitungsvertrags (AVV; oder Data Processing Addendum, DPA) zwingend erforderlich.[38] Dieser Vertrag regelt die Rechte und Pflichten beider Seiten und stellt sicher, dass der Auftragsverarbeiter die Daten nur gemäß den Weisungen des Verantwortlichen verarbeitet.

Hier verläuft die „rote Linie": KI-Anbieter stellen einen solchen AVV ausschließlich für ihre kostenpflichtigen Business-, Team- oder Enterprise-Angebote zur Verfügung.[38]

Die Implikation ist ein klares juristisches Urteil: Der Einsatz von frei verfügbaren Consumer-Diensten (wie ChatGPT Free/Plus, Standard Gemini oder Claude Free/Pro) für jegliche Unternehmensdaten (insbesondere personenbezogene Daten, aber auch Geschäftsgeheimnisse) ist ein direkter Verstoß gegen die DSGVO. Es fehlt die erforderliche Rechtsgrundlage für die Datenverarbeitung.

Die folgende Analyse der Consumer-Tools dient daher primär der Aufklärung über die Risiken, die durch die „Schatten-IT"-Nutzung dieser Tools durch Mitarbeiter entstehen. Die Analyse der Enterprise/API-Produkte bewertet die verbleibenden Risiken, die trotz eines gültigen AVV bestehen.

4.2 Fallstudie: OpenAI (ChatGPT vs. API)

Training

- Consumer (ChatGPT Free/Plus/Pro, Sora): JA. OpenAI verwendet die Inhalte (Prompts und Antworten) von Nutzern dieser Dienste standardmäßig, um seine Modelle zu trainieren und zu verbessern.[41]
- Opt-Out: Nutzer können dieser Verwendung über ein Datenschutzportal widersprechen („do not train on my content").[16]
- Ausnahme: Sogenannte „Temporary Chats" (Provisorische Chats) werden nicht im Verlauf gespeichert und nicht für das Training verwendet.[39]
- Enterprise (API, ChatGPT Enterprise/Team/Edu): NEIN. OpenAI gibt die vertragliche Zusage, dass Kundendaten, die über die API oder die Enterprise-Dienste übermittelt werden, nicht zur Verbesserung oder zum Training der Modelle verwendet werden.[35]

Datensicherheit & Einsicht

- Consumer: JA. Prompts und Konversationen können von autorisierten menschlichen Prüfern bei OpenAI eingesehen werden[44], um die Modellantworten zu validieren oder Feedback zu verarbeiten.[42] Bei Enterprise-Konten (Team/Business) können zudem die Administratoren des eigenen Unternehmens die Chatverläufe ihrer Mitarbeiter einsehen.[16]
- Enterprise (API): BEDINGT JA. Die Daten werden nicht von Menschen zur Modellverbesserung eingesehen. Sie werden jedoch standardmäßig für bis zu 30 Tage aufbewahrt und können automatisiert sowie potenziell menschlich zur „Missbrauchsüberwachung" (Abuse Monitoring) überprüft werden.[35]
- ZDR-Option: Für qualifizierte Enterprise-Kunden bietet OpenAI eine „Zero Data Retention" (ZDR)-Option an. Wenn diese aktiviert ist, werden die Daten nicht mehr für das Abuse Monitoring gespeichert.[35]

Speicherung

- Wo: Die Speicherung der Daten erfolgt primär in den USA.[48]
- Das „OpenAI Ireland"-Problem (Schrems II-Risiko): Viele EU-Unternehmen schließen ihren AVV mit der OpenAI Ireland Ltd. in der Annahme, sich damit sicher im Geltungsbereich der DSGVO zu bewegen. Die Vertragsdokumente (DPA) führen jedoch die US-Muttergesellschaft (OpenAI, LLC) als Unterauftragsverarbeiter (Subprocessor) auf.[38] Dies legitimiert rechtlich einen Datentransfer in die USA. Dort unterliegen die Daten den US-Überwachungsgesetzen (z.B. CLOUD Act), was genau den Sachverhalt darstellt, der im Schrems II-Urteil des Europäischen Gerichtshofs (EuGH) als hochproblematisch eingestuft wurde.[50] Juristische Experten, wie Steiger Legal, bewerten die Rechtslage für

EU-Kunden als „unklar". [33] Im juristischen Kontext bedeutet „unklar" ein nicht kalkulierbares Risiko, dass die Nutzung durch eine Datenschutzbehörde als unzulässig eingestuft wird.

4.3 Fallstudie: Anthropic (Claude Consumer vs. API)

Training

- Consumer (Claude Free/Pro/Max): JA. Mit einer signifikanten Richtlinienänderung im September 2025 werden Nutzerdaten dieser Dienste nun standardmäßig für das Modelltraining verwendet.[52]
- Das aggressive „Opt-Out" und die 5-Jahres-Frist: Das System ist als „Opt-Out" konzipiert. Der Zustimmungsdialog ist jedoch irreführend als „Opt-In" formuliert („You can help improve Claude"), aber standardmäßig vorausgewählt.[53]
- Die Konsequenz: Stimmt der Nutzer zu (auch versehentlich), wird die Aufbewahrungsfrist seiner Daten von den üblichen 30 Tagen auf fünf Jahre verlängert.[53]
- Der Haken: Widerspricht der Nutzer dem Training (Opt-Out), verliert er den Zugriff auf Personalisierungs- und Memory-Funktionen.[56] Diese Kopplung von Dienstleistung an die Einwilligung zur Datenverarbeitung stellt die „Freiwilligkeit" der Einwilligung – eine Kernforderung der DSGVO – stark in Frage.
- Enterprise (API, Claude for Work/Team): NEIN. Diese Dienste sind von der Richtlinienänderung explizit nicht betroffen.[53] Anthropic garantiert vertraglich, API-Daten nicht für das Training zu verwenden.[57]

Datensicherheit & Einsicht

- Consumer: JA. Konversationen, die für eine Sicherheitsüberprüfung markiert werden („flagged for safety review"), werden von Anthropic-Mitarbeitern eingesehen und können zum Training interner Sicherheitsmodelle verwendet werden.[58]
- Enterprise (API): NEIN. Es sei denn, der Nutzer gibt proaktiv Feedback (z.B. über die Daumen-hoch/runter-Funktion). In diesem Fall kann die Konversation zur Qualitätskontrolle verwendet werden.[57]

Speicherung

- Wo: Die Datenspeicherung („Data storage") erfolgt ausschließlich in den USA.[34]
- Der „Multi-Region-Processing"-Haken: Um die Latenz (Geschwindigkeit) für globale Kunden zu verbessern, hat Anthropic im August/September 2025 die Verarbeitung („processing") der Daten auf Rechenzentren in mehreren Regionen, einschließlich Europa und Asien, ausgeweitet.[59] Dies ändert jedoch nichts am Kernproblem für EU-Kunden: Die Daten werden zwar möglicherweise kurz in der EU verarbeitet, aber anschließend permanent in den USA gespeichert.[34] Dies stellt weiterhin einen transatlantischen Datentransfer dar, der unter Schrems II als problematisch gilt.[37]

4.4 Fallstudie: Google (Gemini Apps vs. Vertex AI)

Training

- Consumer (Gemini Apps): JA. Standardmäßig werden die Konversationen und Daten für das Training der KI-Modelle verwendet.[44]
- Der „Privacy vs. Functionality"-Kompromiss: Um das Training zu stoppen, muss der Nutzer die „Gemini Apps-Aktivität" in seinem Google-Konto deaktivieren.[63]
- Der Haken: Das Deaktivieren dieser Einstellung deaktiviert gleichzeitig alle leistungsstarken Erweiterungen (Extensions).[44] Die Integration mit Gmail, Google Docs und Google Drive funktioniert dann nicht mehr.
- Implikation: Google zwingt den Nutzer, wie Anthropic, zu einer Wahl zwischen voller Funktionalität und Datenschutz. Dies ist eine problematische Form der „erzwungenen" Einwilligung im Sinne der DSGVO.
- Enterprise (Vertex AI / Google Cloud): NEIN. Als Teil der Google Cloud Platform bietet Vertex AI eine starke vertragliche „Training Restriction".[66] Weder Prompts noch Kundendaten werden zum Trainieren der Modelle verwendet.[67]

Datensicherheit & Einsicht

- Consumer: JA. Prompts werden von menschlichen Prüfern („human reviewers") bei Google eingesehen, um die Dienste zu verbessern.[44] Google warnt Nutzer explizit davor, vertrauliche Informationen einzugeben.[62]
- Das 3-Jahres-Problem: Von Menschen überprüfte Konversationen werden für bis zu drei Jahre gespeichert, selbst wenn der Nutzer seine Gemini-Aktivität löscht.[69] Google gibt an, diese Daten würden anonymisiert[70], aber die lange Aufbewahrungsfrist stellt ein massives Compliance-Problem dar und widerspricht potenziell dem „Recht auf Vergessenwerden" (Art. 17 DSGVO).
- Enterprise (Vertex AI): NEIN. Prompts und Antworten sind nicht für menschliche Prüfer zugänglich[62], außer zur Missbrauchsüberwachung oder wenn der Kunde explizit Feedback gibt.

Speicherung

- Consumer: Global in Google-Rechenzentren, einschließlich der USA.
- Enterprise (Vertex AI): KLARE KONTROLLE. Dies ist der entscheidende Vorteil der Google-Plattform für EU-Kunden. Als Google Cloud-Dienst bietet Vertex AI klare Data Residency Controls.[30] Ein Unternehmen kann vertraglich und technisch

festlegen, dass seine Daten die EU nicht verlassen.[29] Diese „Data Residency"-Garantie löst das Schrems II-Problem[50] effektiver als die „US-Subprozessor"-Modelle von OpenAI und Anthropic.

4.5 Fallstudie: Meta (Llama und Meta AI)

Llama (Open-Source-Modell)

- Datenschutz: Vollständig vom Hoster abhängig. Wenn ein deutsches Unternehmen Llama 3 auf seinen eigenen Servern in Frankfurt betreibt[17], werden Daten nicht für externes Training verwendet, sind intern gespeichert und werden nur von internen Mitarbeitern eingesehen. Selbst-gehostete Open-Source-Modelle bieten maximale Datenhoheit.[19]
- Acceptable Use Policy: Metas Richtlinie für Llama[72] ist keine Datenschutzrichtlinie. Sie regelt lediglich, wofür das Modell nicht verwendet werden darf (z.B. illegale Aktivitäten, Gewalt, Betrug, Ausbeutung von Kindern).

Meta AI (Consumer-Produkt in Facebook, Instagram etc.)

- Training: JA. Meta trainiert seine KI-Modelle explizit mit den öffentlichen Nutzerdaten seiner Plattformen, einschließlich öffentlicher Posts, Fotos und Bildunterschriften.[74] Private Nachrichten werden (laut Meta) nicht verwendet.[75]
- Opt-Out: Es gibt keinen einfachen Opt-Out-Button.[75] Nutzer (insbesondere in der EU) müssen ein komplexes Formular ausfüllen und detailliert begründen, warum die Verarbeitung sie in ihren Rechten beeinträchtigt. Meta behält sich das Recht vor, diesen Antrag abzulehnen.[77]
- Implikation: Dies ist die datenschutzfeindlichste Implementierung unter den großen Anbietern und für jegliche Unternehmensnutzung völlig unbrauchbar.

5 Synthese und vergleichende Übersicht

Jetzt können wir die folgenden Fragen beantworten.

5.1 Werden meine Daten zur weiteren Entwicklung verwendet?

- Bei Consumer-Produkten (ChatGPT Plus, Claude Pro, Gemini Apps): Ja, standardmäßig. Alle Anbieter nutzen diese Daten für das Training. OpenAI und Google bieten ein Opt-Out an. Das Opt-Out bei Anthropic und Google ist jedoch mit erheblichen Nachteilen verbunden (Funktionsverlust oder 5-Jahres-Speicherung).[44] Bei Meta (Facebook/Instagram) ist ein Opt-Out kaum praktikabel.[77]
- Bei Enterprise/API-Produkten (OpenAI API, Vertex AI, Claude API): Nein. Alle drei großen Anbieter (OpenAI, Anthropic, Google) garantieren in ihren Auftragsverarbeitungsverträgen (AVV), dass Kundendaten aus den API-Diensten nicht für das Training ihrer Modelle verwendet werden.[35]

5.2 Wie sicher sind meine Daten?

- Bei Consumer-Produkten: Geringe Sicherheit. Die Daten sind der Einsichtnahme durch menschliche Prüfer des Anbieters ausgesetzt[44] und werden für potenziell sehr lange Zeiträume gespeichert (bis zu 3 Jahre bei Google[69], 5 Jahre bei Anthropic[54]).
- Bei Enterprise/API-Produkten: Hohe Sicherheit. Die Daten sind durch Branchenstandards (Verschlüsselung, Zugriffskontrollen) geschützt. Das Restrisiko besteht in der temporären Speicherung (meist 30 Tage) für „Abuse Monitoring"[35] und dem potenziellen Zugriff durch US-Behörden (Schrems II-Risiko). Dienste mit „Zero Data Retention" (ZDR)[35] oder garantierter EU-Residenz[29] bieten die höchste Sicherheit.

5.3 Wo werden diese gespeichert?

- OpenAI & Anthropic: Primär in den USA.[33] Selbst wenn ein EU-Vertragspartner (OpenAI Ireland)[38] oder EU-Verarbeitung (Anthropic)[59] involviert ist, bleiben die USA der Speicherort oder ein Unterauftragsverarbeiter, was ein Datentransferrisiko (Schrems II) darstellt.
- Google (Vertex AI): Kontrollierbar. Bietet als einziger der großen US-Anbieter eine klare Option zur Speicherung und Verarbeitung ausschließlich innerhalb der EU.[29]
- Meta (Llama): Beim Self-Hosting ist der Speicherort die eigene Infrastruktur (z.B. Deutschland).[17]

5.4 Wer kann meine Prompts sehen?

Es gibt drei Gruppen, die Prompts einsehen können:

- Anbieter (Menschliche Prüfer): Bei allen Consumer-Produkten.[44] Bei Enterprise/API-Produkten nur für eng definierte Zwecke wie Missbrauchsbekämpfung[35] oder wenn der Nutzer aktiv Feedback gibt.[57]
- Eigene Administratoren: Bei Enterprise-Team-Lizenzen (ChatGPT Team, Claude for Work, Google Workspace) können die Administratoren des eigenen Unternehmens die Konversationen der Mitarbeiter einsehen.[16]
- Öffentlichkeit: Bei der Nutzung von Open-Source-Tools, die von Dritten gehostet werden, oder bei unsicheren Implementierungen.

5.5 Vergleichende Matrix der Daten-governance: Consumer- vs. Enterprise-Modelle (Stand 2025)
Die folgende Tabelle fasst die kritischen Unterschiede zusammen und visualisiert die „rote Linie" (siehe 4.1) zwischen Consumer- und Enterprise-Diensten.

Anbieter	Produkt (Consumer)	Produkt (Enterprise/API)	Standard-Training (Consumer)?	Standard-Training (Enterprise)?	Datenspeicherung (Ort)	Aufbewahrungsfrist (Consumer)	Menschliche Einsicht (Consumer)?	AVV (DPA) verfügbar?
OpenAI	ChatGPT Free/Plus/Pro	API, ChatGPT Enterprise	Ja (Opt-Out möglich)	Nein	USA (mit EU-Subprozessor-Thematik)	30 Tage (nach Löschung	Ja	Nur Enterprise
Anthropic	Claude Free/Pro/Max	API, Claude for Work	Ja (Opt-Out erzwingt 5-Jahres-Frist)	Nein	USA	5 Jahre (bei Training) / 30 Tage (ohne Training)	Ja (für Safety)	Nur Enterprise
Google	Gemini Apps	Vertex AI	Ja (Opt-Out deaktiviert Funktionen	Nein	EU-Residenz möglich	3 Jahre (für überprüfte Chats	Ja	Nur Enterprise
Meta	Meta AI (in Apps)	Llama (Self-Hosted)	Ja (Opt-Out kaum möglich)	Nein (abhängig vom Hoster)	Global / USA	Unbegrenzt	Ja	Nein / N/A

6 Abschließende Empfehlungen für den datenschutzkonformen Einsatz

Basierend auf der vorangegangenen Analyse ergeben sich klare Handlungsempfehlungen für Unternehmen im DACH-Raum, die KI-Modelle datenschutzkonform einsetzen wollen.

6.1 Empfehlung 1: Verbot von Consumer-Tools für Unternehmenszwecke

Die Analyse in Kapitel 4 zeigt unmissverständlich: Die Nutzung von Consumer-Produkten (wie ChatGPT Plus, Claude Pro, Gemini Apps) für jegliche Verarbeitung von Unternehmensdaten (Code, interne Memos, Kundendaten) ist ein klarer Verstoß gegen die DSGVO.

Die Gründe sind:

1. Es ist kein Auftragsverarbeitungsvertrag (AVV) verfügbar.[38]
2. Die Daten werden standardmäßig für das Modelltraining des Anbieters verwendet.[41]
3. Die Daten werden von menschlichen Prüfern des Anbieters eingesehen.[44]
4. Es findet eine unkontrollierte Datenübermittlung in die USA statt.

Unternehmen müssen klare interne Nutzungsrichtlinien erlassen, die den Einsatz dieser „Schatten-IT" verbieten, und dies technisch (z.B. durch Sperrung der Endpunkte) durchsetzen.

6.2 Empfehlung 2: Zwingender Abschluss eines Auftragsverarbeitungsvertrags (AVV)

Für jede Nutzung von externen KI-Diensten (auch Enterprise/API) zur Verarbeitung von Daten, die potenziell personenbezogen sind, ist ein AVV (DPA) nach Art. 28 DSGVO zwingend erforderlich.[38] Dieser Vertrag stellt sicher, dass der Anbieter die Daten nicht für eigene Zwecke (wie Training) verwendet.[35] Ohne gültigen AVV ist der Einsatz illegal.

6.3 Empfehlung 3: Bewertung des transatlantischen Datentransferrisikos (Schrems II)

Das größte verbleibende Risiko trotz eines AVV ist der Datentransfer in die USA. Seit dem „Schrems II"-Urteil des EuGH und dem Ende des „Privacy Shield"-Abkommens[50] sind Datentransfers in die USA, wo Daten dem Zugriff von US-Behörden unterliegen, rechtlich hoch problematisch.

- Das Problem: Sowohl OpenAI (über US-Subprozessoren[33]) als auch Anthropic (durch US-Speicherung[34]) bergen dieses Risiko.
- Die sicherste proprietäre Lösung:

Die Wahl eines Anbieters, der eine vertraglich und technisch garantierte Datenresidenz innerhalb der EU anbietet. Aktuell bietet dies unter den großen US-Anbietern am klarsten Google Cloud mit Vertex AI.[29]

- Die souveränste Lösung (Open-Source): Das Self-Hosting von leistungsfähigen Open-Source-Modellen (z.B. Llama, Mistral) auf eigener Infrastruktur (On-Premise) oder in einer europäischen Private Cloud.[17] Dies eliminiert das Drittanbieter-Datenschutzrisiko vollständig, ist aber mit höheren technischen Kosten verbunden.

6.4 Technische und Organisatorische Maßnahmen (TOMs)

Unabhängig vom gewählten Modell müssen Unternehmen flankierende technische und organisatorische Maßnahmen (TOMs) ergreifen. Dazu gehört die verbindliche Schulung der Mitarbeiter (insbesondere zur „Prompt-Hygiene"), die Implementierung von Anonymisierungs- und Pseudonymisierungs-Gateways, die sensible Daten (wie Namen oder IDs) vor der Übergabe an eine API filtern [78], und die Durchsetzung strenger interner Zugriffskontrollen.

6.5 Zusammenfassende Bewertung

Die Wahl des richtigen KI-Modells im Jahr 2025 ist eine Risikoabwägung. Für unkritische Aufgaben, die keine sensiblen Daten oder Geschäftsgeheimnisse beinhalten, bieten die Enterprise-APIs von OpenAI und Anthropic oft die höchste Leistung.

Sobald jedoch personenbezogene Daten (DSGVO) oder wertvolles geistiges Eigentum (z.B. Quellcode, F&E-Daten) verarbeitet werden sollen, verschiebt sich die Priorität von reiner Leistung zu Compliance. In diesem Szenario sind juristisch nur Lösungen vertretbar, die eine garantierte EU-Datenresidenz bieten (wie Google Vertex AI[29]) oder die das Datenschutzrisiko durch Self-Hosting (Open-Source[17]) vollständig eliminieren.

Quellen und weiterführende Informationen:

[1] Die größten Sprachmodelle (LLMs) - Moin AI, Zugriff am November 3, 2025, https://www.moin.ai/chatbot-lexikon/grosse-sprachmodelle-llms

[2] Which Generative AI is the Best in 2025? GPT-4, Gemini, Claude AI, LLaMA, and More Compared | Pulsebay, Zugriff am November 3, 2025, https://pulsebay.co.nz/post/which-generative-ai-is-the-best-a-deep-dive-into-openai-gemini-claude-ai-meta-and-more/

[3] KI-Modelle 2025 wählen: OpenAI vs. Anthropic vs. Google – Ein praktischer Leitfaden | Fanktank Blog, Zugriff am November 3, 2025, https://www.fanktank.ch/de/blog/choosing-ai-models-openai-anthropic-google-2025

[4] Which AI model should I use with GitHub Copilot?, Zugriff am November 3, 2025, https://github.blog/ai-and-ml/github-copilot/which-ai-model-should-i-use-with-github-copilot/

[5] Comparing GPT, Claude, Llama, and Mistral: Which Large Language Model (LLM) is Right for Your Needs? | Knowledge | Epista, Zugriff am November 3, 2025, https://www.epista.com/knowledge/breakdown-four-of-the-biggest-players-in-ai-gpt-claude-llama-and-mistral

[6] LLM Models: OpenAI ChatGPT, Meta LLaMA, Anthropic Claude, Google Gemini, Mistral AI, and xAI Grok - Mehmet Ozkaya, Zugriff am November 3, 2025, https://mehmetozkaya.medium.com/llm-models-openai-chatgpt-meta-llama-anthropic-claude-google-gemini-mistral-ai-and-xai-grok-bd35779704c2

[7] Popular LLM Models: GPT, Claude, Llama, and Others Compared | by Rizqi Mulki | Medium, Zugriff am November 3, 2025, https://rizqi-mulki.com/popular-llm-models-gpt-claude-llama-and-others-compared-eaed7c5da1b3

[8] Generative KI-Modelle - BSI, Zugriff am November 3, 2025, https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Generative_KI-Modelle.pdf?__blob=publicationFile&v=7

[9] Was ist GitHub Copilot? Ein Expertenleitfaden für 2025 - eesel AI, Zugriff am November 3, 2025, https://www.eesel.ai/de/blog/github-copilot

[10] Die 45 besten KI Tools im Jahr 2025 (in der Praxis getestet) - Synthesia, Zugriff am November 3, 2025, https://www.synthesia.io/de/post/ki-tools

[11] Supported AI models in GitHub Copilot, Zugriff am November 3, 2025, https://docs.github.com/en/copilot/reference/ai-models/supported-models

[12] AI model comparison - GitHub Docs, Zugriff am November 3, 2025, https://docs.github.com/en/copilot/reference/ai-models/model-comparison

[13] LLM vs Generative KI: wie man den Unterschied versteht - Elinext, Zugriff am November 3, 2025, https://www.elinext.de/blog/llm-vs-generative-ai/

[14] Midjourney vs. Stable Diffusion vs. DALL-E 2: Welcher Bildgenerator ist der richtige für dich? (2) - Michael Bickel, Zugriff am November 3, 2025, https://www.michael-bickel.de/2025/01/midjourney-vs-stable-diffusion-vs-dall-e-2-welcher-bildgenerator-ist-der-richtige-fuer-dich-2/

[15] 10 Best AI models you should definitely know about (and why they matter), Zugriff am November 3, 2025, https://pieces.app/blog/best-ai-models

[16] OpenAI Privacy Center, Zugriff am November 3, 2025, https://privacy.openai.com/

[17] Was sind Large Language Models? Und was ist bei der Nutzung von KI-Sprachmodellen zu beachten? - Blog des Fraunhofer IESE, Zugriff am November 3, 2025, https://www.iese.fraunhofer.de/blog/large-language-models-ki-sprachmodelle/

[18] Wie funktionieren LLMs? Ein Blick ins Innere großer Sprachmodelle - Fraunhofer IESE, Zugriff am November 3, 2025, https://www.iese.fraunhofer.de/blog/wie-funktionieren-llms/

[19] Zero Data Retention Policy: The Truth Behind the Promise (and How Your Data Stays Truly Yours) | QuantaMind Blog, Zugriff am November 3, 2025, https://quantamind.co/blog/zero-data-retention-truth-and-how-your-data-stays-yours

[20] Lohnt sich GitHub Copilot? Vorteile, Nachteile und Praxis-Tipps - mindtwo GmbH, Zugriff am November 3, 2025, https://www.mindtwo.de/blog/github-copilot

[21] Visual Studio mit GitHub Copilot – KI-Paarprogrammierung - Microsoft, Zugriff am November 3, 2025, https://visualstudio.microsoft.com/de/github-copilot/

[22] Mit LLMs zur Testfallautomatisierung - Infometis, Zugriff am November 3, 2025, https://www.infometis.ch/blog/mit-llms-chatgpt-zur-testfallautomatisierung-kunstliche-intelligenz-fur-automatisierte-javascript-generierung

[23] KI-gestützte Testfallerstellung - Be | Shaping the Future (DACH), Zugriff am November 3, 2025, https://www.beshapingthefuture.de/insights/ki-gestuetzte-testfallerstellung/

[24] Llama vs. ChatGPT vs. Claude | Best LLM for 2024 - Autonomous, Zugriff am November 3, 2025, https://www.autonomous.ai/ourblog/llama-vs-chatgpt-vs-claude

[25] DALL-E - mindsquare AG, Zugriff am November 3, 2025, https://mindsquare.de/knowhow/dall-e/

[26] Claude 2 - Anthropic, Zugriff am November 3, 2025, https://www.anthropic.com/news/claude-2

[27] Was sind große Sprachmodelle (Large Language Models, LLMs)? | Microsoft Azure, Zugriff am November 3, 2025, https://azure.microsoft.com/de-de/resources/cloud-computing-dictionary/what-are-large-language-models-llms

[28] Claude can now search the web - Anthropic, Zugriff am November 3, 2025, https://www.anthropic.com/news/web-search

[29] Navigating the EU AI Act: Google Cloud's proactive approach, Zugriff am November 3, 2025, https://cloud.google.com/blog/products/identity-security/navigating-the-eu-ai-act-google-clouds-proactive-approach

[30] Data residency | Generative AI on Vertex AI - Google Cloud Documentation, Zugriff am November 3, 2025, https://docs.cloud.google.com/vertex-ai/generative-ai/docs/learn/data-residency

[31] KI-gesteuerte API-Tests in 2025: Vorteile & beste Tools - aqua cloud, Zugriff am November 3, 2025, https://aqua-cloud.io/de/ki-api-ests/

[32] Die Evolution von Low-Code hin zu generativen KI-Modellen wie ChatGPT zur Codegenerierung - VDMA, Zugriff am November 3, 2025, https://vdma.eu/viewer/-/v2article/render/78962507

[33] Gerichtliche Verfügung: OpenAI darf Nutzer-Konversationen mit ..., Zugriff am November 3, 2025, https://steigerlegal.ch/2025/05/18/openai-nutzer-daten-nicht-mehr-loeschen/

[34] Where are your servers located? Do you host your models on EU servers?, Zugriff am November 3, 2025, https://privacy.claude.com/en/articles/7996890-where-are-your-servers-located-do-you-host-your-models-on-eu-servers

[35] Data controls in the OpenAI platform, Zugriff am November 3, 2025, https://platform.openai.com/docs/guides/your-data

[36] How we're responding to The New York Times' data demands in order to protect user privacy | OpenAI, Zugriff am November 3, 2025, https://openai.com/index/response-to-nyt-data-demands/

[37] Anthropic Claude AI In Microsoft 365 Copilot — A Data Boundary Hurdle For The EU?, Zugriff am November 3, 2025, https://ragnarheil.de/anthropic-claude-ai-in-microsoft-365-copilot-a-data-boundary-hurdle-for-the-eu/

[38] ChatGPT und Datenschutz: Was gilt es zu beachten? - Proliance, Zugriff am November 3, 2025, https://www.proliance.ai/blog/chatgpt-datenschutz-welche-folgen-hat-die-nutzung-des-chatbots

[39] Kann ich ChatGPT datenschutzkonform in meinem Unternehmen einsetzen? - eRecht24, Zugriff am November 3, 2025, https://www.e-recht24.de/ki/13409-chatgpt-datenschutz.html

[40] Privacy policy | OpenAI, Zugriff am November 3, 2025, https://openai.com/policies/row-privacy-policy/

[41] Consumer privacy at OpenAI, Zugriff am November 3, 2025, https://openai.com/consumer-privacy/

[42] How your data is used to improve model performance | OpenAI Help ..., Zugriff am November 3, 2025, https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance

[43] Data Usage for Consumer Services FAQ | OpenAI Help Center, Zugriff am November 3, 2025, https://help.openai.com/en/articles/7039943-data-usage-for-consumer-services-faq

[44] Disturbing Privacy Gaps in ChatGPT Plus & Google Gemini Advanced: How to Opt Out & What They're Not Telling You - Reddit, Zugriff am November 3, 2025, https://www.reddit.com/r/Bard/comments/1jywt4x/disturbing_privacy_gaps_in_chatgpt_plus_google/

[45] Data Controls FAQ | OpenAI Help Center, Zugriff am November 3, 2025, https://help.openai.com/en/articles/7730893-data-controls-faq

[46] Privacy policy | OpenAI, Zugriff am November 3, 2025, https://openai.com/en-GB/policies/row-privacy-policy/

[47] Kann mein Chef meine ChatGPT-Anfragen lesen? - Mozilla Foundation, Zugriff am November 3, 2025, https://www.mozillafoundation.org/de/blog/chatgpt-boss-read-privacy/

[48] Data processing addendum | OpenAI, Zugriff am November 3, 2025, https://openai.com/policies/data-processing-addendum/

[49] Zusatz zur Datenverarbeitung - OpenAI, Zugriff am November 3, 2025, https://openai.com/de-DE/policies/data-processing-addendum/

[50] Google Analytics and the GDPR: When will there be legal certainty? - JENTIS, Zugriff am November 3, 2025, https://www.jentis.com/blog/google-analytics-gdpr

[51] Is Google Analytics (3 & 4) GDPR-compliant? [Updated] - Piwik PRO, Zugriff am November 3, 2025, https://piwik.pro/blog/is-google-analytics-gdpr-compliant/

[52] Understanding Anthropic's Data Usage Policy: What Users Need to Know, Zugriff am November 3, 2025, https://rits.shanghai.nyu.edu/ai/understanding-anthropics-data-usage-policy-what-users-need-to-know/

[53] Opt-out statt Opt-in: Anthropic verwendet Nutzerdaten ..., Zugriff am November 3, 2025, https://steigerlegal.ch/2025/09/07/anthropic-claude-ki-training-nutzerdaten/

[54] Updates to Anthropic's Claude AI Terms and Privacy Policy - What You Need to Know, Zugriff am November 3, 2025, https://www.goldfarb.com/updates-to-anthropics-claude-ai-terms-and-privacy-policy-what-you-need-to-know/

[55] Updates to Consumer Terms and Privacy Policy \ Anthropic, Zugriff am November 3, 2025, https://www.anthropic.com/news/updates-to-our-consumer-terms

[56] Anthropic's New Privacy Policy is Systematically Screwing Over Solo Developers - Reddit, Zugriff am November 3, 2025, https://www.reddit.com/r/ClaudeAI/comments/1nd73si/anthropics_new_privacy_policy_is_systematically/

[57] How do you use personal data in model training? | Anthropic Privacy ..., Zugriff am November 3, 2025, https://privacy.claude.com/en/articles/7996885-how-do-you-use-personal-data-in-model-training

[58] Is my data used for model training? | Anthropic Privacy Center - Claude, Zugriff am November 3, 2025, https://privacy.claude.com/en/articles/10023580-is-my-data-used-for-model-training

[59] Anthropic Subprocessor Updates - Trust Center - Anthropic, Zugriff am November 3, 2025, https://trust.anthropic.com/updates

[60] From U.S.-Only to Global: Claude API Regional Processing Launches August 19, 2025, Zugriff am November 3, 2025, https://www.frozenlight.ai/post/frozenlight/694/claude-api-regional-processing-launches-august-19-2025/

[61] Anthropic expanding data processing to include infrastructure in multi regions : r/ClaudeAI, Zugriff am November 3, 2025, https://www.reddit.com/r/ClaudeAI/comments/1m3adw2/anthropic_expanding_data_processing_to_include/

[62] Gemini Apps Privacy Hub - Google Help, Zugriff am November 3, 2025, https://support.google.com/gemini/answer/13594961?hl=en

[63] Zugriff am November 3, 2025, https://support.google.com/gemini/answer/13278892?hl=en&co=-GENIE.Platform%3DAndroid#:~:text=off%20Keep%20Activity%3A-,On%20your%20Android%20phone%20or%20tablet,to%20gemini.google.com.&text=Activity%20.,-Near%20the%20top&text=tap%20Turn%20off-,Turn%20off%20or%20Turn%20off%20and%20delete%20activity.,service%20and%20process%20any%20feedback.

[64] Manage & delete your Gemini Apps activity - Android - Google Help, Zugriff am November 3, 2025, https://support.google.com/gemini/answer/13278892?hl=en&co=GENIE.Platform%3DAndroid

[65] How To Use Gemini Anonymous Mode: Full Step-by-Step Guide - Fello AI, Zugriff am November 3, 2025, https://felloai.com/2025/07/how-to-use-gemini-anonymous-mode-full-step-by-step-guide/

[66] Vertex AI and zero data retention - Google Cloud Documentation, Zugriff am November 3, 2025, https://docs.cloud.google.com/vertex-ai/generative-ai/docs/vertex-ai-zero-data-retention

[67] How Gemini for Google Cloud uses your data, Zugriff am November 3, 2025, https://docs.cloud.google.com/gemini/docs/discover/data-governance

[68] Document AI | Google Cloud, Zugriff am November 3, 2025, https://cloud.google.com/document-ai

[69] Privacy Concerns with Onboard AI: Google Gemini - The University of Tennessee, Knoxville, Zugriff am November 3, 2025, https://oit.utk.edu/security/learning-library/article-archive/privacy-onboard-ai-google-gemini/

[70] Gemini & Datenschutz: Geht das konform? | eRecht24, Zugriff am November 3, 2025, https://www.e-recht24.de/ki/13426-gemini-datenschutz.html

[71] What privacy policy governs the use of AI Studio by a Workspace account? - Reddit, Zugriff am November 3, 2025, https://www.reddit.com/r/googleworkspace/comments/1ll8xd7/what_privacy_policy_governs_the_use_of_ai_studio/

[72] Meta Llama 3 Acceptable Use Policy, Zugriff am November 3, 2025, https://www.llama.com/llama3/use-policy/

[73] meta-llama/Llama-3.1-8B-Instruct - Hugging Face, Zugriff am November 3, 2025, https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[74] Our responsible approach to Meta AI and Meta Llama 3, Zugriff am November 3, 2025, https://ai.meta.com/blog/meta-llama-3-meta-ai-responsibility/

[75] How to opt out of Meta AI: Options to protect your data, Zugriff am November 3, 2025, https://us.norton.com/blog/ai/how-to-opt-out-of-meta-ai

[76] Privacy Matters: Meta's Generative AI Features - About Meta, Zugriff am November 3, 2025, https://about.fb.com/news/2023/09/privacy-matters-metas-generative-ai-features/

[77] Meta's new Privacy Policy: key info for artists - DACS, Zugriff am November 3, 2025, https://www.dacs.org.uk/news-events/what-artists-and-their-beneficiaries-need-to-know-about-metas-new-privacy-policy

[78] Gemini API and Your Data Privacy: A 2025 Guide for Privacy-Conscious Users - Redact, Zugriff am November 3, 2025, https://redact.dev/blog/gemini-api-terms-2025

Abbildung 2: Quelle: Bild generiert von Gemini (Google AI)

Emre Kurtulush ist Consultant bei SEQIS.

Mit einem Hintergrund in Elektrotechnik und Informationstechnik fand er seinen Weg in die Welt der Softwaretests, wo ihn besonders die Verbindung von Technik, Struktur und Qualität begeistert. Erste Erfahrungen sammelte er in technologieintensiven Projekten, bei denen er sein technisches Verständnis und seine analytische Denkweise gezielt einsetzen konnte. Heute liegt sein Fokus auf den technischen Aspekten des Testens – darunter Testautomatisierung, API-Tests und die Optimierung von Testprozessen. Er verfolgt das Ziel, sich kontinuierlich weiterzuentwickeln, neue Werkzeuge und Methoden kennenzulernen und so aktiv zur Verbesserung der Softwarequalität beizutragen.

Denken trotz KI: Warum Automatisierung den Verstand nicht ersetzen darf

von Amine Boutahar

KI-Tools erleichtern unseren Alltag enorm. Doch je mehr wir sie nutzen, desto größer wird die Gefahr, das eigenständige Denken zu verlernen. Der Artikel geht der Frage nach, wie wir kritisch bleiben, wenn Maschinen scheinbar perfekte Antworten liefern.

Denken trotz KI

Planen, entwerfen, argumentieren und umsetzen gelingt mittlerweile präzise und effizient mit KI-Tools wie ChatGPT, Copilot oder Gemini. Was früher Stunden kostete, erledigt die KI in Sekunden. Doch je mehr wir uns auf diese Systeme verlassen, desto stärker wächst die Gefahr, dass wir unser eigenständiges Denken verlernen. Der technologische Fortschritt erleichtert vieles, kann aber auch dazu führen, dass wir unbewusst die Verantwortung für unser Urteilsvermögen abgeben.

Der Artikel geht der Frage nach, wie wir trotz Bequemlichkeit moderner Automatisierung kritisch und selbstständig bleiben können und warum das für die Zukunft der Arbeit entscheidend ist.

Von Entlastung zur geistigen Trägheit

Automatisierung war immer schon ein Synonym für Effizienz. Maschinen haben körperliche Arbeit ersetzt, Software hat Prozesse beschleunigt und nun beginnt KI das Denken selbst zu entlasten.

Texte werden automatisch formuliert, Meetings transkribiert und Lösungswege vorbereitet. Diese Unterstützung spart Zeit, kann aber auch Risiken mit sich bringen. Wir gewöhnen uns an den Komfort, nicht mehr selbst analysieren, hinterfragen oder abwägen zu müssen.

Psychologisch gesehen spricht man hier von kognitiver Trägheit, also der Neigung, auf bestehende Denkmuster oder bequeme Routinen zurückzugreifen, anstatt aktiv nachzudenken. Wenn KI uns plausible Antworten liefert, scheint Nachdenken oft überflüssig. Doch genau hier liegt die Gefahr. Der Verstand, der nicht trainiert wird, verliert seine Schärfe.

Schnelles und langsames Denken

Der Psychologe, Verhaltensforscher und Nobelpreisträger Daniel Kahneman unterscheidet in seinem Buch "Schnelles Denken, langsames Denken" zwischen zwei Denkmodi, die unser Verhalten prägen:

- System 1 ist intuitiv, schnell, automatisch und emotional.
- System 2 ist langsam, reflektiert, analytisch und bewusst.

Künstliche Intelligenz spricht vor allem unser System 1 an. Sie liefert Ergebnisse sofort, elegant formuliert und meist überzeugend. Dadurch aktivieren ihre Antworten unser intuitives Denken, und wir akzeptieren sie oft, ohne sie gründlich zu prüfen.

Das System 2, das für kritische Analyse und tiefes Verständnis verantwortlich ist, bleibt dabei häufig inaktiv. Langsames Denken ist anstrengend, aber essenziell. Es schützt uns vor Fehleinschätzungen und voreiligen Urteilen. Wenn KI unser schnelles Denken verstärkt, müssen wir bewusst Wege finden, das langsame Denken zu fördern.

Viele Menschen neigen jedoch dazu, den Aussagen von automatisierten Systemen blind zu vertrauen. Die KI liefert aber keine absolute Wahrheit, sondern Wahrscheinlichkeiten, die auf Mustern in den zugrunde liegenden Daten basieren. Dieses Vertrauen kann zum Automation Bias führen: Wir übernehmen KI-Empfehlungen oft ungeprüft und vernachlässigen unsere eigenen Einschätzungen, selbst wenn die Vorschläge fehlerhaft oder unvollständig sind.

In der Praxis bedeutet das: Wer sich zu sehr auf KI verlässt, riskiert, dass wichtige Aspekte übersehen oder falsche Entscheidungen getroffen werden. Bewusstes Hinterfragen und kritisches Prüfen bleiben unverzichtbar, um die Stärken von KI effektiv zu nutzen, ohne die eigene Urteilskraft zu verlieren.

Abbildung 1: Kognitive Trägheit durch passive KI-Nutzung
Quelle: Bild generiert von Gemini (Google AI)

Abbildung 2: Aktives Urteilsvermögen mit KI als Werkzeug
Quelle: Bild generiert von Gemini (Google AI)

Kritisches Denken als Schlüsselkompetenz
In einer zunehmend automatisierten Arbeitswelt wird kritisches Denken zu einer zentralen Fähigkeit. Es geht dabei nicht darum, KI zu misstrauen, sondern sie bewusst und reflektiert zu nutzen.

Ein einfaches Vorgehen kann helfen, diese Haltung im Alltag zu verankern.

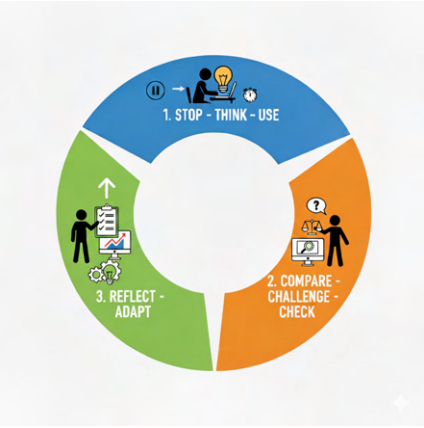


Abbildung 3: Quelle: Bild generiert von Gemini (Google AI)

1. Stop – Think – Use
Bevor man die KI einsetzt, sollte man kurz innehalten. Was will ich wirklich wissen? Welche Art von Antwort brauche ich? Das bewusste Formulieren einer Frage verhindert, dass man die Kontrolle über den Denkprozess abgibt.

2. Compare – Challenge – Check
Nach dem Ergebnis sollte man die Antwort prüfen. Gibt es alternative Quellen oder Sichtweisen? Welche Annahmen stecken in der Antwort? Stimmen die Zahlen, Argumente und Schlussfolgerungen? So bleibt der Mensch im Dialog mit der Maschine und wird nicht zu ihrem passiven Nutzer.

3. Reflect – Adapt
Nach der Nutzung ist Reflexion wichtig. Was hat funktioniert, was nicht? Diese bewusste Nachbetrachtung stärkt langfristig das Urteilsvermögen und verhindert, dass man KI-Ergebnisse unbewusst übernimmt.

Um dieses Vorgehen besser umzusetzen, kann man sogenannte Pre-Prompts verwenden. Ein Pre-Prompt ist eine Art „Meta-Anweisung" oder Verhaltensregel, die man der Aufgabe voranstellt. Man definiert damit die Rolle oder das Ziel der KI-Antwort, bevor man die eigentliche Frage stellt.

Beispiel aus der Softwareentwicklung: „Hilf mir, die Vor- und Nachteile verschiedener Datenstrukturen zu verstehen. Zeige mögliche Denkpfade auf, ohne fertigen Code vorzuschlagen. Ich möchte selbst zur Entscheidung kommen."

Fazit

Automatisierung ist eine enorme Erleichterung, darf aber nicht zu geistiger Abhängigkeit bzw. Verarmung führen. Wenn wir KI nicht als Ersatz, sondern als Werkzeug verstehen und Routinen etablieren, die unser Urteilsvermögen trainieren, kann sie uns helfen, bewusster zu denken. So bleibt der Verstand lebendig, auch in einer zunehmend automatisierten Welt.

Quellen und weiterführende Informationen:

Kahneman, D. (2011). Thinking, Fast and Slow.

https://www.gizmodo.de/kollaborative-intelligenz-wie-man-mentale-traegheit-vermeidet-und-die-formel-mensch-ki-staerkt-2000027399

https://cset.georgetown.edu/publication/ai-safety-and-automation-bias/

https://biases.de/automation-bias/

https://stealthesethoughts.com/2025/07/22/stop-ai-hijacking-your-thinking/

https://www.resilienz-akademie.com/resilienz-allgemein/metakognition/

https://www.forbes.com/sites/charlestowersclark/2025/03/14/how-ai-changes-critical-thinking-new-microsoft-research-findings/

Amine Boutahar ist Agile Development Wizard.

Als Full Stack Developer bei SEQIS widmet er sich besonders der Weiterentwicklung des SEQITrackers. Dabei legt er seinen Fokus auf die Gestaltung ansprechender UI-Designs.

Sein Credo: Vom Backend bis hin zum Frontend qualitativ, sicher, funktional und benutzerfreundlich!

Referenzstory: Otto Bock Healthcare Products GmbH

mit Manius Schinkel

ottobock.

„Mit SEQIS haben wir einen verlässlichen Partner gefunden, mit dem wir seit langem zusammenarbeiten und gemeinsam hochwertige Produkte erfolgreich umsetzen."

Manius Schinkel, Head of Engineering Digital Health Applications bei Ottobock

Otto Bock Healthcare Products GmbH

Der globale MedTech-Champion Ottobock vereint über 100-jährige Tradition mit herausragender Innovationskraft in den Bereichen Prothetik, Neuro-Orthetik und Exoskelette. Ottobock entwickelt innovative Versorgungslösungen für Menschen mit eingeschränkter Mobilität und treibt die Digitalisierung der Branche voran.

Gegründet 1919 in Berlin, ist das Unternehmen mit rund 9.300 Mitarbeitenden heute an 45 Standorten weltweit aktiv und betreibt das größte Patientenversorgungsnetzwerk mit rund 400 Standorten weltweit.

Die Mission, die Bewegungsfreiheit, Lebensqualität und Unabhängigkeit von Menschen zu verbessern, ist tief in der DNA des Unternehmens verwurzelt.

Die Aufgabenstellung

- Entwicklungsdruck meistern: Parallele Entwicklung mehrerer mobiler Apps für medizinische Geräte
- Effizienz schaffen
- Einhaltung der strengen regulatorischen Anforderungen
- Möglichkeit zur Skalierung

Wie wurde die Aufgabenstellung gelöst

- Aufbau einer zukunftssicheren Testinfrastruktur als Fundament
- Teststrategie: Mix aus manuellen Tests und Testautomatisierung
- Professionelle Unterstützung im Testzyklus: Von der Spezifikation bis zur lückenlosen Dokumentation
- Partnerschaftliche Zusammenarbeit: Expertise, Flexibilität und Vertrauen ermöglichen schnelle Ressourcen Skalierung

Das Gespräch

- Manius Schinkel, Head of Engineering Digital
- Sandra Benseler, SEQIS Sales Managerin

Effizienz und Sicherheit in der Medizintechnik: Eine Erfolgsgeschichte über Vertrauen und Qualität

In stark regulierten Branchen ist **Qualitätssicherung** keine Option, sondern das Fundament des Erfolgs. Als Ottobock, das führende Unternehmen im Bereich der Medizintechnik, vor der Herausforderung stand, die Softwareentwicklung um mobile Anwendungen zu erweitern, war klar: Es braucht einen Partner, der nicht nur Testaufgaben übernimmt, sondern strategisch mitdenkt. Wir haben mit Herrn Schinkel, Head of Engineering, über unsere Zusammenarbeit gesprochen.

Sandra Benseler: Herr Schinkel, danke, dass Sie sich die Zeit nehmen. Ihr Unternehmen stand vor der Herausforderung, in einem hochregulierten Markt mehrere **mobile Anwendungen** zu entwickeln. Was war die ursprüngliche Ausgangslage?

Manius Schinkel: Zu Beginn haben wir erkannt, dass uns die interne Erfahrung für die parallele Entwicklung mehrerer mobiler Apps fehlte. Da Effizienz und ein strukturiertes Vorgehen für uns entscheidend waren, haben wir uns gezielt externe Unterstützung gesucht.

Sandra Benseler: Wie genau hat SEQIS Sie dabei unterstützt, diese anspruchsvollen Ziele zu erreichen?

Manius Schinkel: Die Zusammenarbeit geht weit über reines Testing hinaus. SEQIS hat für uns die **Infrastruktur für die Testautomatisierung** aufgebaut und gleichzeitig das **manuelle Testen** übernommen. Entscheidend ist jedoch das tiefe Verständnis

für unsere Prozesse. SEQIS Expert:innen haben sich so intensiv in unsere komplexe Dokumentation eingearbeitet, dass sie uns schon bei der **Anforderungsanalyse** wertvollen Input geben konnten. Sie kennen unsere Produkte mittlerweile so gut, dass sie für unser Team eine wertvolle Unterstützung sind.

Sandra Benseler: Welche besonderen Herausforderungen bestehen in Ihren Projekten, da Ihre Branche stark reguliert ist?

Manius Schinkel: In unserem Sektor, der Medizintechnik, sind Qualität und Sicherheit nach Normen wie ISO 13485 absolut entscheidend. Dazu kamen neue Herausforderungen wie **Cybersecurity und Barrierefreiheit**. Wir mussten also schnell eine professionelle Testinfrastruktur aufbauen.

Sandra Benseler: In dynamischen Projekten ist Flexibilität ein Schlüsselfaktor. Welche Rolle spielt die Skalierbarkeit der SEQIS Unterstützung für Sie?

Manius Schinkel: Eine sehr große. Die **Flexibilität**, die SEQIS bietet, ist ein enormer Vorteil. Wenn wir für ein großes Release kurzfristig mehr Ressourcen benötigen, erweitern wir die Zusammenarbeit unkompliziert. Das ist deutlich schneller und agiler, als intern neue Mitarbeiter:innen einzustellen. Diese **Skalierbarkeit** gibt uns eine enorme Planungssicherheit. Wir wissen, dass wir auf Lastspitzen und Änderungen im Projektverlauf sofort reagieren können.

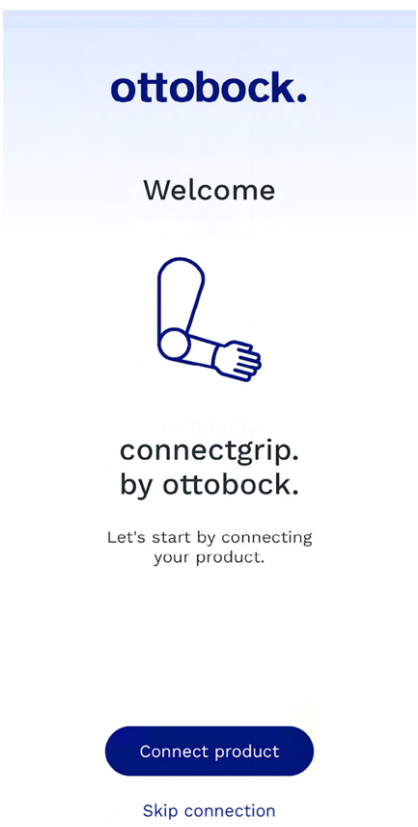
Sandra Benseler: Eine solche Partnerschaft ist auch ein Investment. Wie hat sich die Zusammenarbeit wirtschaftlich für Sie ausgezahlt?

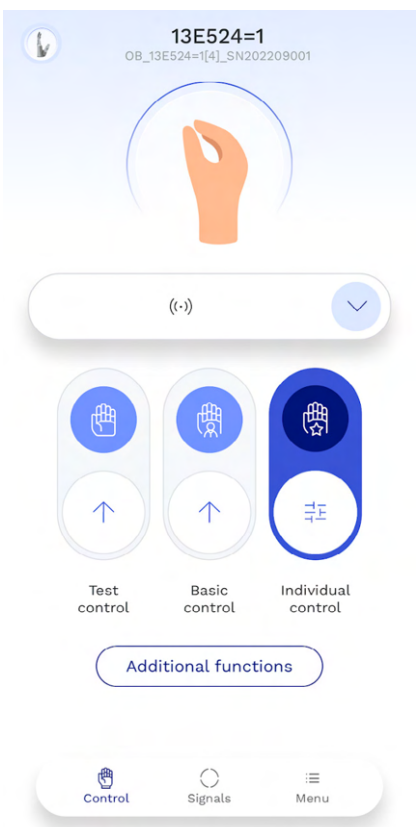
Manius Schinkel: Das Investment hat sich schnell amortisiert. Mit den erfahrenen Expert:innen von SEQIS und der Möglichkeit auch kurzfristig zu skalieren, sparen wir uns lange Einarbeitungszeiten. Die etablierte Testautomatisierung steigert unsere Effizienz bei jedem neuen Release. Das Feedback aus allen Abteilungen war durchwegs positiv, weil die **hohe Qualität der Arbeit** und die **partnerschaftliche Zusammenarbeit** uns einfach überzeugt haben.

Dank der Fachkenntnisse von SEQIS und der **flexiblen Ressourcenskalierung** machte sich die Zusammenarbeit schnell bezahlt. Die anfängliche Einarbeitungsphase wurde durch die schnelle Anpassungsfähigkeit und Kapazitätsbereitstellung mehr als ausgeglichen, was eine dynamische Projektanpassung ermöglicht.

Sandra Benseler: Was würden Sie abschließend als den größten Mehrwert der Zusammenarbeit bezeichnen und was raten Sie anderen Unternehmen in einer ähnlichen Situation?

Manius Schinkel: Der größte Mehrwert ist das **vertrauensvolle Verhältnis**. Wir müssen unsere Partner nicht permanent kontrollieren, weil wir wissen, dass die Qualität stimmt. Mein Rat ist klar: Behandeln Sie Qualitätssicherung nicht stiefmütterlich. Indem man sich Expert:innen wie SEQIS ins Haus holt, gewinnt man nicht nur fachliches Know-how, sondern auch einen unschätzbaren **Blick von außen**, der hilft, die eigenen Prozesse zu optimieren. Das ist keine reine Dienstleistung, das ist eine strategische Partnerschaft, die auf Vertrauen basiert und den Erfolg sichert.





Quelle: https://play.google.com/store/apps/details?id=com.ottobock.connectgrip&utm_source=e-mea_Med&pli=1

Lokales AI-Hosting auf dem Mac: Ein praktischer Leitfaden für Coder

von Daniel Kleissl

Nachdem ich in meinem letzten Artikel meine Erfahrungen mit dem Cloud-basierten AI-Assistenten Codeium geteilt habe, hier der nächste Schritt: der Umstieg auf ein vollständig lokales Setup. Die Gründe dafür liegen auf der Hand: maximale Kontrolle über sensible Daten, keine Latenz durch Internetverbindungen, die Möglichkeit zur Offline-Nutzung und die Freiheit, mit den verschiedensten Open-Source-Modellen zu experimentieren. Für jeden, der einen Mac mit Apple Silicon nutzt, ist die gute Nachricht: Dank der Metal-Architektur und cleverer Software wie llama.cpp ist das Hosting von leistungsstarken Modellen nicht nur möglich, sondern erstaunlich performant.

In diesem Artikel werde ich unsere Erkenntnisse zum Hosting von zwei Coding-Modellen teilen: Qwen3 Coder 32b und Devstral. Ich werde die empfohlenen Kommandozeilenparameter aufschlüsseln, die wichtigsten Sampling-Parameter praxisnah erläutern und aufzeigen, wie man jedes Modell für seine spezifischen Stärken optimal konfiguriert.

Das Fundament: llama.cpp und die Magie der Quantisierung

Für das lokale Hosting auf einem Mac ist llama.cpp das mit Abstand wichtigste Werkzeug. Es ist eine in C/C++ geschriebene Inferenz-Engine, die für Apple Silicon optimiert ist und die GPU (via Metal) für die rechenintensiven Matrizenmultiplikationen nutzt.

Der Schlüssel zur Lauffähigkeit großer Modelle auf Consumer-Hardware liegt in der Quantisierung. Ein Modell wie Qwen3 Coder 32b hätte ursprünglich mit 16-Bit-Gleitkommazahlen (FP16) eine Größe von über 64 GB. Die Quantisierung reduziert die Präzision der Gewichte im neuronalen Netz, z.B. auf 4-Bit-Ganzzahlen. Modelle für Llama.cpp kommen im GGUF-Format. Bei der Auswahl einer Datei stößt man auf Bezeichnungen wie Q4_K_M oder Q5_K_S. Für den Einstieg hat sich Q4_K_M als exzellenter Kompromiss aus Performance, geringer Dateigröße und hoher Qualität erwiesen.

Wenn es die Hardware erlaubt, gilt hier allerdings: Größer (Q8 statt Q4) ist eindeutig besser, wenn es um Genauigkeit der Antwort geht. Ist eine schnelle Antwort Priorität, dann empfiehlt es sich, die niedrigste Quantisierung zu wählen, die noch gute Antworten liefert. Im Normalfall bedeutet das alles inklusive Q4. Es ist auf jeden Fall ein wenig Experimentierfreudigkeit gefragt, um das optimale Modell und die besten Parameter für den eigenen Anwendungsfall herauszufinden.

> Wer mehr zu Quantisierung und die technischen Hintergründe davon erfahren möchte, findet unter diesem Link einen Startpunkt für die eigene Recherche: https://huggingface.co/docs/optimum/en/concept_guides/quantization

Die Beispiele im hinteren Teil des Artikels sollen hier als Startpunkt dienen. Will man hier größere Anpassungen vornehmen, sollte man sich als ersten Schritt genauer über die Hintergründe und Auswirkungen der einzelnen Parameter informieren. Der Detailgrad und die Komplexität können hier schnell ins Unermessliche wachsen. Und selbst wenn man sich für neue Parameter entschieden hat, dann wollen diese am besten auch getestet werden. Da allerdings Änderungen in den Parametern mitunter subtile oder schwer quantifizierbare Änderungen im Verhalten des Modells verursachen, ist Vorsicht gefragt.

Im Normalfall findet man für jedes Modell im Web genügend Vorschläge dazu, und wir haben uns ebenfalls an diesen orientiert, um den Testaufwand klein zu halten. Meistens steht der Aufwand für den umfassenden Test dieser Parameter in keiner Relation zu den damit nachweisbar erreichten Verbesserungen.

Die Kunst des Samplings: Die wichtigsten Parameter im Überblick
Bevor wir in die modellspezifischen Konfigurationen eintauchen, hier eine kurze Übersicht der wichtigsten Sampling-Parameter, die das Verhalten der AI steuern und deren Bedeutung:

Abbildung 1: Quelle: Bild generiert von Gemini (Google AI)

Parameter	Zweck	Auswirkung
--temp <wert>	Kreativität: Regelt die Zufälligkeit der Wortwahl.	Niedrig (~0.2): Deterministisch, präzise. Hoch (~0.9): Kreativ, aber fehleranfälliger.
--top-k <wert>	Fokus: Beschränkt die Auswahl auf die K wahrscheinlichsten Wörter.	Verhindert unsinnige Wörter, kann aber zu repetitiven Antworten führen.
--top-p <wert>	Dynamischer Fokus: Wählt Wörter, deren summierte Wahrscheinlichkeit p erreicht.	Flexibler als Top-K, oft ein besserer Kompromiss zwischen Kreativität und Kohärenz.
--mirostat <modus>	Adaptive Steuerung: Versucht, die „Überraschung" (Perplexität) konstant zu halten.	Alternative zu temp/top-p, gut für lange, kohärente Texte. 2 ist ein gängiger Modus. Taucht selten in empfohlenen Parametern auf, sollte daher als experimentell betrachtet werden.
--repeat-penalty <wert>	Anti-Wiederholung: Bestraft das Modell für das Wiederholen von Wörtern.	Ein Wert von 1.1 ist fast immer eine gute Idee, um Endlosschleifen zu vermeiden.

Beispiele aus der Praxis

Modell 1: Qwen3 Coder 32b – Der logische Problemlöser
Qwen3 Coder ist ein Allrounder mit einer besonderen Stärke im logischen Denken und der Generierung von qualitativ hochwertigem, korrektem Code. Mit seinem 32k-Token-Kontextfenster kann es problemlos größere Dateien verarbeiten. Es eignet sich hervorragend für Aufgaben, bei denen es auf algorithmische Korrektheit und die Einhaltung komplexer Anweisungen ankommt.

Anwendungsfall A: Präzise Code-Generierung und Unit-Tests
Hier wollen wir, dass das Modell exakt unseren Anweisungen folgt und vorhersagbaren Code generiert. Kreativität ist unerwünscht.

Empfohlene Parameter:

```
./llama-server -m Qwen3-32B-Coder-Q4_K_M.gguf
-ngl 99
-c 32768 #kurzform von --ctx-size
-t -1 #-1 setzt Aanzahl erlaubter Threads auf unbeschränkt
--temp 0.2
--top-k 10
--repeat-penalty 1.1
```

Warum diese Parameter?

- -ngl 99 oder --n-gpu-layers 99: Stellt sicher dass alle Layer des Modells in den GPU-Speicher geladen werden
- --temp 0.2: Eine sehr niedrige Temperatur zwingt das Modell, fast immer das statistisch wahrscheinlichste nächste Token zu wählen. Das Ergebnis ist deterministischer und weniger fehleranfällig – perfekt für das Ausformulieren einer bekannten Logik.
- --top-k 10: Wir schränken die Auswahl auf die 10 wahrscheinlichsten Tokens ein. Dies eliminiert das Risiko, dass das Modell auf eine abwegige, aber „kreative" Idee kommt, die den Code unbrauchbar machen würde.
- Anwendungsbeispiel: Schreibe eine Python-Funktion, die einen Bubble-Sort-Algorithmus implementiert. Füge Type-Hints und einen Docstring hinzu, der die Zeitkomplexität erklärt.

Anwendungsfall B: Algorithmisches Brainstorming und Refactoring-Ideen
Manchmal wissen wir nicht genau, wie die Lösung aussehen soll, und möchten, dass das Modell verschiedene Ansätze vorschlägt.

Empfohlene Parameter:

```
./llama-server -m Qwen3-32B-Coder-Q4_K_M.gguf
-ngl 99
-c 32768
-t -1
--temp 0.7
--top-p 0.95
--repeat-penalty 1.1
```

Alternativ mit Mirostat:

```
./llama-server -m Qwen3-32B-Coder-Q4_K_M.gguf
-ngl 99
-c 32768
-t 1
--mirostat 2
--repeat-penalty 1.1
```

Warum diese Parameter?

- --temp 0.7 & --top-p 0.95: Diese Kombination erlaubt dem Modell, auch weniger offensichtliche Tokens in Betracht zu ziehen, solange sie zur Gruppe der 95% wahrscheinlichsten Kandidaten gehören. Dies fördert die Generierung alternativer Lösungswege, ohne komplett unsinnig zu werden.

- --mirostat 2: Mirostat ist eine exzellente Alternative, da es versucht, den Output über längere Passagen hinweg interessant zu halten, was ideal für das Ausformulieren verschiedener Konzepte ist.

- Anwendungsbeispiel: Ich habe eine langsame Funktion zur Datenverarbeitung in Pandas. Zeige mir drei verschiedene Wege, wie ich sie optimieren könnte, z.B. durch Vektorisierung oder die Verwendung von Dask.

Modell 2: Devstral – Der Architekt mit dem Super-Gedächtnis
Devstrals herausragendes Merkmal ist sein gigantisches 128k-Token-Kontextfenster. Das entspricht etwa 100 Seiten Text. Damit kann das Modell nicht nur eine einzelne Datei, sondern auch eine ganze Codebasis „im Kopf behalten". Seine Stärke liegt in groß angelegten Refactorings, der Analyse von Projektarchitekturen sowie in Aufgaben, die Konsistenz über viele Dateien hinweg erfordern.

Hinweis: Die Nutzung des vollen Kontextfensters erfordert eine beträchtliche Menge an RAM (idealerweise 64 GB oder mehr).

Seine wahre Stärke spielt Devstral in Kombination mit dem Agent-Framework „Open-Hands" aus. Darauf näher einzugehen würde allerdings den Rahmen dieses Artikels sprengen. Wer sich selbst dazu schlau machen will oder das Framework ausprobieren möchte, finden alle Infos dazu hier: https://docs.openhands.dev/

Anwendungsfall A: Großflächiges und konsistentes Refactoring
Stellen Sie sich vor, Sie müssen in einem ganzen Projekt eine veraltete Bibliothek ersetzen oder von JavaScript auf TypeScript migrieren. Devstral kann durch seinen großen Kontext gut Zusammenhänge zwischen verschiedenen Files tracken und dadurch besser bei solchen "hochvernetzten" Tasks unterstützen.

Wichtig: Umso größer der Kontext, umso länger dauert auch die Antwort. Je nach Hardware kann das schon mal mehrere Minuten dauern. Deshalb bei so komplexen Tasks am besten die Logs im Auge behalten, um sicherzustellen, dass das Modell noch arbeitet und nicht "hängt".

Für unsere Tests haben wir das auf Huggingface verfügbare quantisierte Modell von der Gruppe "Unsloth" verwendet.

WICHTIGER HINWEIS: Zum Zeitpunkt der Veröffentlichung dieses Artikels wies die neueste "2507"-Version von Unsloth einen Bug im Zusammenhang mit --jinja auf. Daher verwenden wir das "2505"-Modell, das diesen Bug nicht hat.

Empfohlene Parameter: (von Unsloth)

```
./llama-server -m devstral-Q4_K_M.gguf
--threads -1
--ctx-size 131072
--cache-type-k q8_0
--n-gpu-layers 99
--seed 3407
--prio 2
--temp 0.15
--repeat-penalty 1.15 #höher als bei Unsloth
--min-p 0.01
--top-k 64
--top-p 0.95
--jinja
```

Warum diese Parameter?

- --ctx-size 131072: Der wichtigste Parameter hier. Wir geben dem Modell den maximalen Kontext, damit es Abhängigkeiten zwischen Dateien versteht.
- --temp 0.15: Bei einem so großen Refactoring ist Konsistenz entscheidend. Eine extrem niedrige Temperatur stellt sicher, dass das Modell den einmal gewählten Stil und die Namenskonventionen beibehält.
- --repeat-penalty 1.15: Leicht erhöht, um bei sehr langen Generierungen über mehrere Dateien hinweg subtile Wiederholungen im Stil oder in Kommentaren zu vermeiden.
- --jinja: Aktiviert die Verwendung des Chat-Templates in der Jinja-Syntax, das mit dem Modelfile automatisch mitgeliefert wird

- Anwendungsbeispiel: Hier sind fünf miteinander verbundene Klassen. Refaktoriere sie, um das Singleton-Pattern durch Dependency Injection zu ersetzen. Achte darauf, alle Instanziierungen im gesamten Code konsistent zu aktualisieren.

Anwendungsfall B: Codebase-Analyse und Onboarding für neue Entwickler
Ein neues Teammitglied muss sich in ein komplexes Projekt einarbeiten. Devstral kann als persönlicher Projekt-Tutor fungieren.

Empfohlene Parameter:

```
./llama-server -m devstral-Q4_K_M.gguf
--threads  -1
--ctx-size 131072
--cache-type-k q8_0
--n-gpu-layers 99
--seed 3407
--prio 2
--temp 0.5
--repeat-penalty 1.15 #höher als bei Unsloth
--min-p 0.01
--top-k 64
--top-p 0.9
--jinja
```

Abbildung 2: Quelle: Bild generiert von Gemini (Google AI)

Warum diese Parameter?

- --temp 0.5: Wir benötigen hier eine verständliche, menschenähnliche Erklärung, keine exakte Code-Generierung. Eine mittlere Temperatur erzeugt einen flüssigeren, natürlicheren Sprachstil.

- --top-p 0.9: Gibt dem Modell genügend Flexibilität, um komplexe Zusammenhänge auf verschiedene Weisen zu formulieren.

- Anwendungsbeispiel: Basierend auf der gesamten Codebasis, die ich dir gebe: Erkläre die Hauptverantwortlichkeiten des ‚AuthenticationService' und wie er mit dem ‚User' und ‚Session' Modell interagiert. Erstelle ein Mermaid.js-Diagramm, das diese Beziehungen visualisiert.

Fazit

Der Schritt zum lokalen Hosting von AI-Modellen auf meinem Mac war eine unglaublich lehrreiche Erfahrung. Die präzise Kontrolle über die Sampling-Parameter, abgestimmt auf die Stärken spezifischer Modelle wie Qwen3 und Devstral, war der Schlüssel um diese Modelle auf "Consumer-Hardware" gehosted unseren Developern zur Verfügung stellen zu können

Somit bleiben alle mit der AI geteilten Informationen in unserer Kontrolle und es erlaubte uns, AI-unterstütztes Coding auch in hochsensiblen Projekten einsetzen zu können. Und alle gesammelten Learnings auf unserem Weg zur selbstgehosteten AI machen wir über razzfazz.ai für alle verfügbar.

Quellen und weiterführende Informationen:

https://huggingface.co/mistralai/Devstral-Small-2507

https://huggingface.co/unsloth/Devstral-Small-2505-GGUF

https://docs.unsloth.ai/models/tutorials-how-to-fine-tune-and-run-llms/devstral-how-to-run-and-fine-tune

https://github.com/ggml-org/llama.cpp/tree/master/tools/server

https://huggingface.co/

https://github.com/ggml-org/llama.cpp/tree/master

https://docs.openhands.dev/

https://razzfazz.ai/

https://mistral.ai/news/devstral

https://qwen.ai/home

Dieser Artikel wurde mit Unterstützung von AI verfasst

Daniel Kleissl ist Agile Development Wizard.

In der Weiterentwicklung des SEQITrackers konnte er wertvolle Erfahrung im Bereich Full Stack Development sammeln und sein Wissen aus seinem Fachhochschulstudium vertiefen. Zusätzlich konnte er sich durch die Jira-Cloud-Plugin Entwicklung wichtiges Know-How im Cloud-Development aufbauen.

Für ihn sind Clean Code und eine gute Testabdeckung die wichtigsten Säulen eines guten Entwicklungsprozesses.

SEQIS remotetesting.net

SEQIS Remote Testing Services - Hilfe aus Österreich für Ihre spezifischen Testaufgaben.

Kontaktieren Sie mich unter alexander.weichselberger@SEQIS.com und sprechen wir über Ihre Aufgabenstellungen. Wir helfen Ihnen, Ihr Projekt zum Erfolg zu führen.

QR Code scannen und weitere Informationen erhalten.

5 Fragen an die KI-Experten: Wie künstliche Intelligenz die Unternehmenswelt revolutioniert

von Sandra Benseler

Künstliche Intelligenz (KI) ist längst mehr als nur ein futuristisches Konzept – sie ist ein entscheidender Wettbewerbsfaktor in nahezu jedem Geschäftsbereich. Von der Softwareentwicklung bis zum Vertragsmanagement: Unternehmen suchen nach Wegen, KI effizient und sicher zu integrieren. Wir haben fünf zentrale Fragen an KI-Experten gestellt, die Aufschluss darüber geben, wie Unternehmen die neue Technologie am besten nutzen können.

1. Wie setze ich KI am besten für die Testfallerstellung ein?

KI revolutioniert das Software-Testing durch die Automatisierung und Optimierung der Testfallerstellung. Der beste Ansatz ist, KI dort einzusetzen, wo sie ihre Stärken voll ausspielen kann:

- Generierung von Testfällen: KI-gestützte Tools können modellbasiert aus bestehender Code-Struktur oder Spezifikationsdokumenten Testfälle generieren. Sie erkennen dabei auch Grenzfälle und komplexe Szenarien
- Testabdeckung: KI kann automatisch neue und unvorhersehbare Testeingaben erzeugen, sowie durch die Analyse großer Mengen von Testdaten die Testabdeckung signifikant erhöhen.
- Dynamische Priorisierung: KI-Systeme können Testfälle dynamisch nach Relevanz und kritischen Funktionen priorisieren, um sicherzustellen, dass die wichtigsten Tests zuerst ausgeführt werden.

Wichtig: Das Ziel ist nicht, menschliche Tester zu ersetzen, sondern sie von repetitiven Aufgaben zu entlasten, damit sie sich auf komplexes, kreatives Problem-Lösen konzentrieren können.

2. Kann mich die KI beim Schreiben von Code unterstützen?

Ja, KI-Coding-Tools sind mittlerweile zu einem Must-have für Entwickler geworden und steigern die Produktivität massiv. Aktuelle KI-Modelle dienen als vielseitige Programmierassistenten:

- Code-Vervollständigung und -Generierung: Tools wie GitHub Copilot oder spezialisierte KI-Agenten können in Echtzeit Code-Zeilen vorschlagen, vervollständigen oder ganze Code-Blöcke auf Basis von einfachen Anweisungen generieren. Sie interpretieren den gesamten Projektkontext, um kohärente Lösungen zu liefern.
- Fehlerbehebung und Analyse: Die Systeme helfen, Sicherheitslücken schnell zu erkennen, Fehler aufzuspüren und Lösungsvorschläge zu liefern.
- Automatisierung von Routinen: KI kann Unit-Tests und Dokumentationen automatisch generieren.

Wichtig: Trotz der Geschwindigkeit und des Komforts ist menschliche Überprüfung unerlässlich. Der von der KI generierte Code muss immer auf Korrektheit, Effizienz und Sicherheitslücken geprüft werden.

3. Wie verhindere ich, dass (sensible) Daten außerhalb des Unternehmens oder meiner Abteilung landen?

Der Schutz sensibler und personenbezogener Daten (DSGVO-Konformität) ist eine der größten Herausforderungen beim Einsatz von KI-Tools, insbesondere bei cloudbasierten Diensten. Die Experten geben hier klare Handlungsempfehlungen:

- Verwendung geschlossener Systeme (On-Premise oder Private Cloud): Für sensible Daten sollten Unternehmen interne oder private KI-Lösungen in Betracht ziehen, die sicherstellen, dass Daten die Unternehmensgrenzen nicht verlassen.
- Daten-Anonymisierung und Pseudonymisierung: Bevor Daten in externe KI-Tools (z.B. große Sprachmodelle wie ChatGPT oder Gemini in der Standardversion) eingegeben werden, müssen sie anonymisiert werden. Vertrauliche Unternehmensdaten, proprietärer Quellcode oder personenbezogene Kundendaten dürfen niemals in frei zugängliche, öffentliche KI-Chatbots eingegeben werden, da sie dort potenziell zum Training des Modells verwendet werden können.
- Vertragliche Regelungen: Bei der Nutzung von Enterprise-Lösungen sollte ein Auftragsverarbeitungsvertrag (AVV) abgeschlossen werden, der die Nutzung der Daten für Trainingszwecke explizit ausschließt.
- Mitarbeiter-Richtlinien: Klare interne Nutzungsrichtlinien und Schulungen sind essenziell, um Mitarbeiter für die Risiken beim Umgang mit sensiblen Daten in KI-Systemen zu sensibilisieren.

4. Wie finden wir die optimale KI-Lösung für unsere spezifische geschäftliche Herausforderung und welche Auswahlkriterien sind dabei entscheidend?

Die Wahl des richtigen KI-Tools erfordert einen strukturierten Ansatz, der über generische Anwendungsfälle hinausgeht. Es beginnt mit einer klaren Zieldefinition: Was genau soll die KI leisten (z.B. Texterstellung, Datenanalyse, komplexe logische Schlussfolgerungen)? Für einfache, repetitive Aufgaben genügen oft schlankere Modelle, während höchste Präzision und tiefes Verständnis der Daten leistungsstarke Reasoning-Modelle (wie fortschrittliche LLMs) erfordern. Die entscheidenden Auswahlkriterien sind:

1. Integrationsfähigkeit: Lässt sich die KI nahtlos in unsere bestehende IT-Infrastruktur (CRM, ERP, etc.) einbetten?

2. Skalierbarkeit und Kosten: Wächst das Tool mit unseren Anforderungen und ist das Preismodell transparent?

3. Datensouveränität und Sicherheit: Bietet der Anbieter On-Premise- oder Private-Cloud-Lösungen an, um die Kontrolle über unsere sensiblen Daten zu behalten?

4. Usability und Anpassung: Ist das Tool für die Mitarbeiter intuitiv bedienbar und können wir es auf unternehmensspezifische Prozesse und Fachsprachen trainieren?

5. Wie können KI-Systeme unseren Mitarbeitern helfen, schnell und zuverlässig die gesuchten Informationen in unserem internen „Daten-Silo" – also Verträgen, E-Mails oder alten Projektakten – zu finden?
Der Einsatz von wissensbasierten KI-Systemen transformiert das interne Wissensmanagement. Diese Lösungen basieren oft auf Technologien wie Retrieval-Augmented Generation (RAG) und agieren als intelligente, unternehmensweite Suchmaschine für Ihre unstrukturierten Daten (Texte, Dokumente).

- Präzise Antworten statt Dokumentenflut: Mitarbeiter müssen nicht mehr Hunderte von Dokumenten durchsehen. Die KI analysiert den gesamten Datenbestand (Verträge, E-Mail-Archive, Akten) und generiert direkte, präzise Antworten auf komplexe Fragen. Sie liefert nicht nur eine Liste von Suchergebnissen, sondern fasst die relevanten Informationen zusammen und verweist auf die Originalquelle.

- Beschleunigte Routineaufgaben: Im Rechtswesen kann die KI in Sekunden relevante Klauseln, Risiken oder Fristen in Verträgen identifizieren. Im Kundenservice ermöglicht sie den schnellen Zugriff auf Produktwissen aus alten Projektakten.

- Fazit: Dies führt zu einer drastischen Reduzierung der Suchzeit und stellt sicher, dass Entscheidungen auf der Grundlage des gesamten, aktuellen Unternehmenswissens getroffen werden können, anstatt nur auf leicht verfügbare Informationen zurückzugreifen.

Die Antworten der Experten zeigen: Künstliche Intelligenz ist kein optionales Extra mehr, sondern ein strategischer Wettbewerbsfaktor. Die Potenziale sind enorm, insbesondere dort, wo KI repetitive, zeitintensive Aufgaben übernimmt und menschliche Expertise erweitert.

Allerdings wird der Erfolg von KI-Initiativen entscheidend davon abhängen, ob Unternehmen die Balance zwischen Innovation und Sicherheit meistern. Der Schutz sensibler Daten hat oberste Priorität: Nur durch den bewussten Einsatz von geschlossenen (On-Premise) Systemen, konsequenter Anonymisierung und klaren Compliance-Richtlinien kann das Risiko des Datenabflusses verhindert werden.

Kurz gesagt: Die Zukunft gehört jenen Unternehmen, die nicht nur die richtige KI für die richtige Aufgabe auswählen, sondern auch die notwendigen organisatorischen und technischen Schutzmauern errichten, um die Vorteile der KI sicher und verantwortungsvoll zu nutzen. KI steigert die Produktivität massiv – aber nur, wenn sie mit strategischer Weitsicht und klarem Fokus auf Datensouveränität implementiert wird.

Der Text wurde mit Unterstützung des KI-Modells Gemini (Google) erstellt und redaktionell überarbeitet.

Sandra Benseler ist Sales Managerin bei SEQIS.

Als primäre Kundenbetreuerin ist sie die zentrale Ansprechpartnerin für sämtliche Kundenbelange. Sie begleitet Kunden umfassend von der ersten Anfrage über die gesamte Projektlaufzeit bis zum erfolgreichen Abschluss – und gewährleistet so eine durchgehende, verlässliche Betreuung.

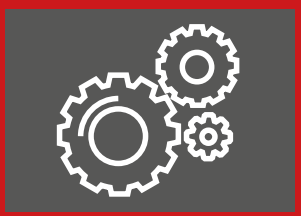
Sandra Benseler ist Ihre direkte und erreichbare Ansprechpartnerin für alle Kundenanliegen und Fragen zum gesamten Service-Portfolio von SEQIS und razzfazz.

SEQIS ist der führende österreichische Anbieter in den Spezialbereichen
IT Analyse, Development, Softwaretest und Projektmanagement.
Beratung, Verstärkung und Ausbildung:
Ihr Partner für hochwertige IT-Qualitätssicherung.

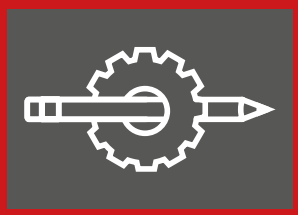
IT ANALYSE
Notwendige Änderungen analysieren und IT-gerecht aufbereiten

DEVELOPMENT
Agil, individuell und qualitätsgesichert

TESTING
Probleme durch methodischen Soll-Ist Vergleich erkennen



RELEASE & OPERATE
Reibungsloser Go Live und Betrieb der IT-Lösungen



DEVOPS
Neuerungen abgestimmt mit Entwicklung und Betrieb live setzen

METHODOLOGY & TOOLS
Vorgehensweisen optimieren und auf die richtigen Tools setzen

TRAINING & WORKSHOPS
Mitarbeiter Know-how stärken - standardisiert oder maßgeschneidert

PROJEKT-MANAGEMENT
verantwortlich, zielorientiert und pragmatisch

Analysis. Development. Test. Management. Better Results.
www.SEQIS.com