TRAIBER.NRW
TRANSFORMATION
DER AUTOMOTIVE INDUSTRIE
IN DER BERGISCHEN REGION

Responsible AI in Unternehmen

Gefördert durch:

Bundesministerium
für Wirtschaft
und Energie

aufgrund eines Beschlusses
des Deutschen Bundestages

TRAIBER.NRW
TRANSFORMATION
DER AUTOMOTIVE INDUSTRIE
IN DER BERGISCHEN REGION

Responsible AI in Unternehmen

veröffentlicht im Rahmen des Projektes TRAIBER.NRW

PROJEKTINFORMATION

Die fortschreitende Digitalisierung, die Mobilitätswende und der notwendige Wandel zur Klimaneutralität fordern die Automobilzulieferer und die Bergische Region an vielen Stellen. Das Bundesministerium für Wirtschaft und Energie fördert TRAIBER.NRW im Rahmen der Förderbekanntmachung „Transformationsstrategien für Regionen der Fahrzeug- und Zulieferindustrie“ mit 4,1 Mio. EUR bis Ende 2025.

Die Bergische Region umfasst die Städte Remscheid, Solingen, Wuppertal und Düsseldorf sowie den Kreis Mettmann, den Rhein-Kreis-Neuss, den Ennepe-Ruhr-Kreis und den Oberbergischen Kreis. Projektpartner sind automotiveland.nrw e.V., die Bergische Universität Wuppertal, die Gemeinschaftslehrwerkstatt der Industrie von Velbert und Umgebung e.V. (GLW), die Heinrich-Heine-Universität Düsseldorf und die Hochschule Bochum. Das Projekt wurde von den Sozialpartnern der Region initiiert und wird von ihnen maßgeblich unterstützt und begleitet: VBU® Vereinigung Bergischer Unternehmerverbände e.V., Arbeitgeberverband Remscheid und Bergisches Land e.V., IG Metall Velbert, IG Metall Ennepe-Ruhr-Wupper und IG Metall Remscheid-Solingen.

www.traiber.nrw

Inhaltsverzeichnis

1 Responsible AI im Unternehmenskontext

Künstliche Intelligenz (KI) hat in den vergangenen Jahren den Schritt aus Forschung und Pilotprojekten in den Unternehmensalltag vollzogen. Spätestens mit dem Erfolg generativer KI-Anwendungen ist sichtbar geworden, dass KI nicht mehr nur ein Spezialthema der Informatik ist, sondern Geschäftsprozesse, Produkte und Dienstleistungen in der Breite prägt. Ob in der Qualitätssicherung, in der Produktionsplanung, im Kundenservice oder in der Verwaltung: KI-basierte Systeme unterstützen Entscheidungen, automatisieren Abläufe und erschließen neue Geschäftsmöglichkeiten, gerade dort, wo sie direkt in operative und industrielle Prozesse eingebettet ist. Aktuelle Studien zeigen, dass ein Großteil der deutschen Unternehmen KI inzwischen als geschäftskritisch betrachtet und entsprechende Budgets aufstockt[1]. Für Organisationen aller Größen, insbesondere kleine und mittelständige Unternehmen (KMU), stellt sich damit nicht mehr die Frage, ob KI eingesetzt wird, sondern wie dies verantwortungsvoll und dauerhaft tragfähig gelingt, also rechtssicher, erklärbar bzw. nachvollziehbar, ressourcenschonend und datensouverän, ohne Risiken für Sicherheit, Qualität, Compliance und Reputation zu unterschätzen.

Dieses Themenpapier führt in das Konzept von „Responsible AI“ ein und richtet sich an Verantwortliche in Unternehmen, die KI einsetzen oder einführen möchten, mit besonderem Fokus auf kleine und mittlere Unternehmen (KMU). Ziel ist es, Orientierung zu bieten, wie KI so eingeführt und weiterentwickelt werden kann, dass sie Wert schafft, gleichzeitig Risiken beherrschbar bleiben und als Teil einer verantwortungsvollen Digitalisierung in die Geschäftsprozesse integriert werden kann, jenseits kurzfristiger Trends und mit Blick auf nachhaltige Wertschöpfung.

Das Themenpapier ist so aufgebaut, dass es eine schnelle Orientierung ermöglicht. Technische Details werden nur dort vertieft, wo sie für Entscheidungen und Umsetzung relevant sind. Zu jedem Themenfeld gibt es eine verständliche Einordnung sowie Hinweise, welche Kapitel für welche Bereiche besonders wichtig sind. Die Schwerpunkte sind:

- der **EU AI Act** als rechtlicher Rahmen für den KI-Einsatz in Europa,
- **Explainable AI** als Grundlage für Nachvollziehbarkeit und Vertrauen,
- **Green AI** als Ansatz für ressourcenschonende und wirtschaftliche KI,
- **lokale Modelle (On-Premise/Edge)** als Grundlage für datensouveränen, kontrollierbaren Betrieb und strategische Unabhängigkeit.

Das Themenpapier ist modular aufgebaut: Jedes Kapitel beginnt mit einer kurzen Einordnung und vertieft anschließend die für Umsetzung und Governance relevanten Aspekte. Für einen **schnellen Einstieg** ordnet die folgende Übersicht die Kapitel typischen Verantwortungsbereichen zu:

Kapitel 2 – EU AI Act: Leitung/Management, Recht/Compliance, Datenschutz.

Nutzen: Einordnung von Pflichten, Rollen und Risikoklassen. Grundlage für Governance, Dokumentations- und Umsetzungsbedarf im KI-Einsatz.

Kapitel 3 – Explainable AI: Data Science/Analytics, IT, Fachbereichsleitungen.

Nutzen: Orientierung, welche Formen von Erklärbarkeit in der Praxis benötigt werden. Brücke zwischen Modelllogik, fachlicher Nachvollziehbarkeit und Vertrauen in Entscheidungen.

Kapitel 4 – Green AI: ESG (Environmental, Social & Governance), IT, Controlling.

Nutzen: Ansatzpunkte zur Reduktion von Energie- und Ressourcenverbrauch. Verbindung von Umweltwirkung, Kostenperspektive und technischer Umsetzung im KI-Betrieb.

Kapitel 5 – Lokale Modelle (On-Premise/Edge): IT/Architektur, Informationssicherheit

Nutzen: Entscheidungsgrundlage zur Betriebsform (lokal vs. Cloud). Kriterien für Datensouveränität, Sicherheitsanforderungen und Infrastrukturfolgen.

Kapitel 6 – Zusammenfassung & Ausblick: Leitung/Management, Recht/Compliance.

Nutzen: Verdichtung der Kernaussagen. Konsolidierte Orientierung für nächste Schritte, Prioritäten und Schnittstellen zwischen den Verantwortungsbereichen.

2 Der EU AI Act

Der *EU AI Act* (Regulation (EU) 2024/1689) ist das weltweit erste umfassende Gesetz zur Regulierung von Künstlicher Intelligenz. Der AI Act folgt einem *risikobasierten Ansatz*: Die Pflichten richten sich danach, wie riskant ein bestimmter KI-Einsatz ist. Ziel der Regulierung ist es, die Potenziale von KI verantwortungsvoll zu nutzen, ohne dabei Menschenrechte, Unternehmensrechte oder geltende Gesetze zu verletzen. Internationale Leitlinien wie die OECD Principles on AI und die UNESCO Recommendation on the Ethics of Artificial Intelligence bilden hierfür die normative Grundlage und betonen eine menschenzentrierte, transparente und diskriminierungsfreie Gestaltung von KI-Systemen. Für kleine und mittlere Unternehmen (KMU) ergeben sich daraus neue Chancen, aber auch Unsicherheiten und Mehraufwände, da viele Aufgaben nicht an eine Rechtsabteilung ausgelagert werden können. Der Einsatz von KI wird damit zur Managementaufgabe, die Strategie und Rechtssicherheit zusammenführt, einschließlich möglicher Haftungsfragen. Je nachdem, ob Unternehmen KI-Systeme selbst entwickeln, einkaufen oder lediglich einsetzen, gelten unterschiedliche Pflichten und Nachweisanforderungen, insbesondere bei Hochrisiko-KI-Anwendungen, für die der AI Act unter anderem Risikobewertungen, Dokumentations-, Transparenz- und Aufsichtspflichten vorsieht.

2.1 Für wen gilt der AI Act?

Privatpersonen, die KI nur für rein persönliche Zwecke nutzen, sind vom EU-AI-Act ausgenommen. Für Unternehmen aber gilt: **Wer KI beruflich einsetzt, fällt unter das Gesetz**. Schon ehrenamtliche Tätigkeiten im Verein können dazu führen, dass die Schwelle zur geschäftlichen Nutzung überschritten ist. Der EU-AI-Act gilt hierbei bereits unabhängig davon ob er bereits vollständig in nationales Recht überführt wurde oder nicht.

Zentral im Sinne des EU-AI-Acts ist dabei die eigene *Rolle* eines Unternehmens: Sind Sie Anbieter (Provider) von KI z.B. als Betriebsmittel oder sind Sie Betreiber (Deployer)?

- **Anbieter (Provider):** Entwickeln oder modifizieren Sie eine KI-Anwendung und bieten diese unter eigenem Namen/Marke an? Dann sind Sie Provider – mit umfangreichen Pflichten. *Beispiel:* Sie nutzen ein Open-Source-KI-Modell wie Llama von Meta und bauen damit ein internes Tool, das Sie unter Ihrem Firmennamen dem Team oder Kundinnen und Kunden bereitstellen. Auch wenn Sie eine bestehende KI-Software in ein eigenes Produkt „einkapseln“ und anbieten, gelten Sie als Anbieter.
- **Betreiber (Deployer):** Kaufen oder mieten Sie KI-Software (z. B. einen Chatbot oder eine Automatisierungssoftware) ein und nutzen diese in Ihrem Unternehmen, ohne das System selbst zu verändern oder unter eigenem Namen zu vermarkten? Dann sind Sie Betreiber. *Beispiel:* Sie

lizenzieren einen bestehenden KI-Service für Ihre Webseite oder zur Dokumentenklassifikation – in dem Fall fallen Sie unter die Betreiberpflichten.

Ein typischer Grenzfall in der Praxis ist der **lokale Betrieb eines KI-Modells (Self-Hosting)**. Wenn Sie z. B. ein Large Language Model (LLM) auf eigenen Servern betreiben und Ihren Beschäftigten zur Verfügung stellen, ohne eigene wesentliche Anpassungen und ohne das System extern zu vertreiben, sind Sie in der Regel ebenfalls als Anbieter zu betrachten. Solange Sie keine Hochrisiko-KI einsetzen, bleibt der regulatorische Aufwand dabei meist überschaubar: keine Registrierung, keine technischen Dokumentationspflichten, insbesondere Transparenz (z. B. klarstellen, dass die Interaktion mit einer KI erfolgt).

2.2 Welche KI-Systeme betrifft das Gesetz?

Das meiste im AI Act dreht sich um die *Risikoklasse* der jeweiligen KI-Anwendung. Es gibt drei Hauptgruppen:

1. **Verbotene Praktiken:** KI-Systeme, die z. B. Menschen manipulieren oder ohne Einwilligung biometrisch überwachen, sind untersagt.
2. **Hochrisiko-KI:** Hierzu zählen KI-Systeme in sensiblen Bereichen wie Medizin, Justiz, Personalrekrutierung, kritische Infrastruktur oder Bildung (z. B. automatisierte Leistungsbewertung).
3. **Sonstige (geringes Risiko):** Die meisten Anwendungen im KMU-Alltag, etwa Produkt-Empfehlungen im Webshop, Chatbots für Standardfragen, Rechtschreibkorrektur etc. fallen unter die Kategorie „geringes Risiko“. Für sie gelten keine Registrierungspflichten, keine technischen Dokumentationspflichten – lediglich ggf. Transparenz (Hinweis, wenn Nutzende mit einer KI interagieren).

2.3 Was müssen Unternehmen konkret tun?

Der EU AI Act ergänzt bestehende Regelwerke und gilt *ohne Vorgriff* auf datenschutzrechtliche Vorgaben. Sobald KI-Systeme personenbezogene Daten verarbeiten, bleibt die DSGVO uneingeschränkt anwendbar. Das betrifft insbesondere automatisierte Entscheidungen mit erheblichen Auswirkungen (Art. 22 DSGVO) sowie Transparenz, Zweckbindung, Datenminimierung und Privacy-by-Design.

Zur Einordnung der Rechtsgrundlagen ist aktuell Bewegung im System: Mit dem Digital Omnibus schlägt die Europäische Kommission gezielte Klarstellungen zur Anwendung des berechtigten Interesses (Art. 6 Abs. 1 lit. f DSGVO) im Kontext von KI-Systemen und KI-Modellen vor, insbesondere entlang des KI-Lebenszyklus (z. B. Training, Testen, Validieren)[2]. In eine ähnliche Richtung weist auch die Rechtsprechung: Das OLG Köln hat im Eilverfahren den Unterlassungsantrag gegen die Nutzung öffentlich gestellter Profildaten zum KI-Training zurückgewiesen und die Verarbeitung

bei summarischer Prüfung grundsätzlich als auf Art. 6 Abs. 1 lit. f DSGVO stützbar eingeordnet[3]. Bei externen KI-Services entscheidet der *Vertragstyp* maßgeblich über Datenschutzrisiken: Bei Consumer-Angeboten (z. B. ChatGPT Free/Plus) kann die Nutzung von Inhalten zur Modellverbesserung per Opt-out unterbunden werden, während Business-Angebote (z. B. API, ChatGPT Enterprise/Business) nach Anbieterangaben standardmäßig nicht zum Training genutzt werden und eine Datenschutzvereinbarung (DPA) unterstützen[4], [5]. Für den Unternehmenseinsatz sollten Verträge daher mindestens festlegen: Rollen und Verantwortlichkeiten (inkl. Art. 28 DSGVO, falls Auftragsverarbeitung), keine Trainingsnutzung ohne Opt-in, Regeln zu Aufbewahrung und Löschung sowie Transparenz zu Unterauftragnehmern und Datenübermittlungen.

Ergänzend zur DSGVO kommen durch den EU AI Act weitere, risikobasierte Pflichten hinzu. Für KMU sind typischerweise die ersten drei Punkte immer relevant. Zusätzliche Anforderungen ergeben sich je nach Rolle (Anbieter/Betreiber) und Risikoklasse:

- **Risikoprüfung und Dokumentation:** Zunächst ist zu klären: *Welche Rolle* haben wir? Ist unser System Hochrisiko? Die Abgrenzung ist nicht immer trivial – viele Begriffe sind juristisch noch nicht abschließend geklärt. Für Hochrisiko-Systeme sind eine technische Dokumentation, Risikomanagement und ggf. eine Registrierung bei Behörden erforderlich.
- **Transparenzpflichten:** Nutzende müssen erkennen können, wenn sie mit einer KI interagieren (außer, es ist ohnehin offensichtlich, z. B. animierte Spielfigur). Wer synthetische Medien (Deepfakes, KI-generierte Bilder/Videos/Texte mit Täuschungspotenzial) erstellt und veröffentlicht, muss dies kennzeichnen – wobei es Spielräume gibt, wie sichtbar das erfolgen muss. Für künstlerische oder satirische Werke reicht ein allgemeiner Hinweis im Kontext.
- **AI Literacy:** Anbieter und Betreiber von KI-Systemen sollen nach besten Kräften Maßnahmen ergreifen, um ein ausreichendes Maß an AI Literacy bei den Personen sicherzustellen, die KI-Systeme in ihrem Auftrag betreiben oder nutzen. Was „ausreichend“ bedeutet, ist bewusst kontext- und risikobasiert; die EU-Kommission betont hierzu ausdrücklich Flexibilität und verzichtet aktuell auf starre Mindestvorgaben. In der Praxis bietet sich eine Baseline-Awareness für alle Beschäftigten an, die KI-gestützte Funktionen im Arbeitsalltag nutzen (z. B. Office-Copilot, interne Chatbots, Videokonferenzsysteme mit automatischer Transkription oder Protokollerstellung). Für risikoreichere Anwendungen, insbesondere Hochrisiko-Systeme, sind rollenbezogene Vertiefungen sinnvoll; die Kommission weist zudem darauf hin, dass bei Hochrisiko-Systemen zusätzliche Anforderungen an Schulung und menschliche Aufsicht greifen können.
- **General-Purpose AI und Open Source:** Wer eigene KI-Modelle (insbesondere große, generative Modelle) entwickelt und anbietet, muss zusätzliche Transparenz über Trainingsdaten, Urheberrechtsstrategie und ggf. Energieverbrauch schaffen. Diese Pflichten können eigentlich nur in der Anbieter-Rolle entstehen. Open-Source-Modelle sind in einigen Punkten erleichtert, aber auch hier gelten bestimmte Pflichten.

2.4 Typische Herausforderungen für Unternehmen

Auch wenn viele Anforderungen des EU AI Act klar formuliert sind, entsteht in der Umsetzung häufig zusätzlicher Klärungsbedarf. Unternehmen sehen sich dabei sowohl juristischen als auch organisatorischen und technischen Herausforderungen gegenüber. Besonders relevant sind die folgenden Punkte.

- **Rechtsunsicherheit:** Viele Definitionen (z. B. wann ist ein System Hochrisiko?) werden erst durch zukünftige Praxis und Rechtsprechung klarer.
- **Dokumentationsaufwand:** Im Hochrisikobereich ist der Aufwand beträchtlich; im Minimal-Risiko-Bereich bleibt es meist überschaubar.
- **Haftung:** Wer ein KI-System unter eigenem Namen anbietet, trägt die Produktverantwortung. Einfache Rebranding-Lösungen (API-Wrapper mit eigenem Logo) können schnell dazu führen, dass man als „Anbieter" mit allen Pflichten gilt. Generell gilt: Das Unternehmen haftet (in der Regel) vollumfänglich für den Einsatz einer KI, so wie es in einer gewerblichen Unternehmung auch der Fall ist. Der Einsatz von KI reduziert in diesem Kontext die Haftung nicht.
- **Technische Umsetzung:** Kennzeichnungspflichten (Wasserzeichen, Metadaten) sind nicht immer mit vorhandenen Tools sicher umsetzbar. Der Stand der Technik ist dynamisch, und es gibt Grauzonen.

2.5 Fazit

Der AI Act ist keine unüberwindbare Hürde, verlangt aber bewusste Entscheidungen, sorgfältige Rollenklärung und gegebenenfalls professionelle Beratung – besonders beim Einsatz oder Angebot von Hochrisiko-KI. Letztendlich haftet ein Unternehmen für die Produkte des Unternehmens und seine Erstellung nach wie vor – unabhängig davon ob KI eingesetzt wird oder nicht. Wer einfache KI-Tools nutzt, kann mit überschaubaren Aufwänden und klaren Regeln planen. Innovation wird nicht verhindert, aber die Spielräume für einen Einsatz von KI hängen stark davon ab, wie die eigene Rolle und der Einsatzzweck konkret definiert sind.

3 Explainable Artificial Intelligence

Viele moderne KI-Modelle liefern beeindruckend gute Ergebnisse, sind aber von außen kaum nachvollziehbar. Für Unternehmen entsteht dadurch ein Spannungsfeld: Einerseits sollen Entscheidungen effizient automatisiert werden, andererseits müssen sie gegenüber Kunden, Mitarbeitenden und Aufsichtsbehörden erklärbar sein. *Explainable Artificial Intelligence* (XAI, deutsch: erklärbare künstliche Intelligenz) verfolgt das Ziel, KI-Entscheidungen transparenter zu machen und Begründungen zu liefern, die Menschen verstehen können. Dieses Kapitel zeigt, warum Erklärbarkeit nicht nur ein technisches, sondern auch ein geschäftskritisches Thema ist.

Explainable AI umfasst also Methoden und Techniken, die sichtbar machen, wie ein KI-System zu einer Entscheidung gelangt. Dabei geht es nicht nur um technische Details, sondern auch um Antworten auf die Frage „Kann ein Mensch nachvollziehen, warum die KI genau dieses Ergebnis liefert?“. XAI ermöglicht es KI einzusetzen ohne mit einer undurchsichtigen Black-Box agieren zu müssen [6], deren Mechanismen und Vorhersagen weder erkennbar noch überprüfbar sind. Man kann auch sagen, dass Explainable AI, der Versuch eines Brückenschlags zwischen unverständlichen maschinellen Entscheidungsprozessen und dem menschlichen Bedürfnis nach Erklär- und Nachvollziehbarkeit ist. Es werden drei zentrale Begriffe unterschieden:

- **Transparenz** bezeichnet die nachvollziehbare Dokumentation und Offenlegung relevanter Informationen über ein KI-System entlang seines Lebenszyklus. Dazu gehören je nach Kontext u. a. Zweck und Grenzen des Systems, Leistungskennzahlen, sowie Informationen zur Herkunft und Struktur der verwendeten Daten und zum Training, Validieren und Testen (z. B. Datenbeschreibung, Datenaufbereitung, Evaluationssetup)[7]–[10].
- **Interpretierbarkeit** meint, dass ein Modell für einen Menschen aus seiner Struktur heraus verständlich ist, ohne nachträglich angewendete Approximationen oder Erklärungsmethoden zu benötigen. Es soll direkt nachvollziehbar sein, wie eine Vorhersage zustande kommt[11], [12].
- **Erklärbarkeit** beschreibt Verfahren, die das Verhalten oder eine konkrete Entscheidung eines Modells für einen bestimmten Zweck verständlich machen. Dies umfasst unterschiedliche Formen (global vs. lokal) und Techniken (z. B. Attributionsmethoden, Gegenfaktisches, Surrogate)[11], [13].

In den letzten zehn Jahren hat sich Explainable AI von einem Forschungsthema zu einem praktisch unverzichtbaren Element beim KI-Einsatz in Unternehmen entwickelt. An erklärbarer KI führt kaum ein Weg vorbei, wenn Verantwortlichkeit, Vertrauen und Transparenz gewährleistet werden sollen. Überall dort, wo KI entscheidende Prozesse mitbestimmt, brauchen wir Nachvollziehbarkeit. Insbesondere Transparenz wird explizit durch den EU AI Act eingefordert (siehe Kapitel 2) und wird

damit zum unumgänglichen Thema für den Einsatz von KI-Systemen. Erklärbarkeit (Explainability) ermöglicht es Unternehmen darüber hinaus, die „Black-Box KI“ zu öffnen und so KI vertrauenswürdiger und effektiver zu machen. Dies schließt nahtlos an Prinzipien der Responsible AI an. Erklärbare KI schafft Transparenz über die Funktionsweise von Modellen und hilft, Bias/Fairness-Probleme aufzudecken, bevor sie zum realen Schaden führen. Bias und Fairness Probleme entstehen, wenn ein KI-System bestimmte Gruppen aufgrund fehlerhafter oder unausgewogener Daten systematisch anders behandelt als andere.

3.1 Warum ist Explainable AI für Unternehmen relevant?

Für Unternehmen ist das Thema aus mehreren Gründen geschäftsrelevant. KI wird zunehmend in Bereichen eingesetzt, in denen Entscheidungen direkte wirtschaftliche Konsequenzen haben, etwa bei der Qualitätskontrolle, der Ressourcenplanung, der Wartungsprognose oder im Kundenservice. Ohne Erklärbarkeit (Explainability) bleibt unklar, ob ein Modell zuverlässig arbeitet, ob es systematische Fehler aufweist oder ob es zu Entscheidungen kommt, die gegen interne Vorgaben oder gesetzliche Rahmenbedingungen verstoßen (siehe Kapitel 2). Fehlt diese Nachvollziehbarkeit, wird KI schnell als unsicher oder riskant wahrgenommen, was Innovationen ausbremst, Verantwortlichkeiten verwischt und die Akzeptanz für den Einsatz von KI signifikant beschädigen kann. Wer einmal gesehen hat, wie ein KI-System einen Fehler macht, welcher nicht nachvollziehbar ist, der verliert Vertrauen.

Mit XAI lassen sich solche Risiken kontrollieren. Anwender und Entwickler können überprüfen, ob ein Modell stabile und plausible Muster nutzt, ob einzelne Faktoren übermäßig starken Einfluss haben oder ob Datenverzerrungen zu unfairen oder ineffizienten Ergebnissen führen. Für Unternehmen bedeutet das: bessere Entscheidungsgrundlagen, geringere Haftungsrisiken und ein höheres Vertrauen in KI-gestützte Prozesse. Gleichzeitig erleichtert Explainable AI die Kommunikation zwischen Fachabteilungen und Management, da Erklärungen typischerweise in visuelle oder sprachlich gut verständliche Form gebracht werden [14].

In der Unternehmenspraxis wird die Bedeutung von Explainable AI besonders deutlich, wenn man konkrete Anwendungsszenarien betrachtet. Die folgenden Beispiele zeigen typische Situationen, in denen Erklärbarkeit direkt über Vertrauen, Akzeptanz und die Qualität von Entscheidungen mit KI-Unterstützung entscheidet. Sie können als Orientierung dienen, um eigene Anwendungsfälle im Unternehmen zu identifizieren, bei denen XAI einen besonderen Mehrwert bietet.

- KI-basierte Qualitätsprüfung, bei der nachvollziehbar sein soll, warum ein Bauteil als Ausschuss klassifiziert wurde, um Reklamationen, Audits und interne Qualitätsdiskussionen fundiert führen zu können.
- Empfehlungssysteme für Produkte oder Dienstleistungen, deren Vorschläge für Vertrieb und Kunden verständlich begründet werden sollen, damit Entscheidungen nicht „blind“ übernommen werden.

- KI-gestützte Priorisierung von Service-Tickets oder Beschwerden, bei der Mitarbeitende verstehen müssen, warum bestimmte Fälle bevorzugt bearbeitet werden, um Fairness und Kundenzufriedenheit sicherzustellen.
- Unterstützung von Sachbearbeitung durch KI, z. B. bei Schadensmeldungen, Kreditanträgen oder Prüfung von Dokumenten, wo Entscheidungen begründbar sein müssen, weil sie direkte finanzielle oder rechtliche Folgen haben.
- KI-gestützte Bewerbervorauswahl, bei der transparent sein muss, warum bestimmte Kandidatinnen und Kandidaten bevorzugt werden, um Diskriminierung zu vermeiden und gegenüber Betriebsrat und Bewerbenden argumentieren zu können.
- Interne Berichte und Dashboards, bei denen Kennzahlen oder Prognosen aus KI-Modellen stammen und gegenüber Management und Fachbereichen so erklärt werden müssen, dass Entscheidungen darauf gestützt werden können.

3.2 Welche Arten von Explainable AI gibt es in der Praxis?

XAI umfasst unterschiedliche Methoden und Perspektiven. In der Praxis unterscheidet man hauptsächlich:

1. **Modellinterne Erklärbarkeit (Intrinsic Interpretability):** Hierbei handelt es sich um Modelle, deren Aufbau bereits von Natur aus nachvollziehbar ist. Ihre Struktur macht sichtbar, wie einzelne Eingaben zu einer Entscheidung führen, etwa durch klare Entscheidungsregeln oder überschaubare Rechenwege. Die Erklärbarkeit entsteht also direkt aus dem Modell selbst, ohne dass zusätzliche Analysen notwendig sind [15].
2. **Modellunabhängige Erklärbarkeit (Post-Hoc):** Bei diesem Ansatz werden Erklärungen erst nach der eigentlichen Modellentscheidung erzeugt. Das bedeutet: Das Modell selbst bleibt unverändert, aber sein Verhalten wird im Nachhinein analysiert, um nachvollziehbare Hinweise auf Einflussfaktoren, Entscheidungslogiken oder mögliche Alternativen zu geben. Solche Verfahren können zum Beispiel zeigen, welche Merkmale für eine bestimmte Vorhersage besonders wichtig waren oder wie sich die Entscheidung verändern würde, wenn einzelne Eingaben anders gewesen wären. Da diese Methoden unabhängig von der internen Struktur funktionieren, eignen sie sich auch für sehr komplexe, schwer einsehbare Modelle [16].
3. **Datenbezogene Erklärbarkeit:** Hier steht nicht das Modell, sondern der Datensatz im Mittelpunkt. Ziel ist es, zu verstehen, wie die Zusammensetzung der Daten das Modellverhalten beeinflusst. Dazu gehört beispielsweise, Ungleichgewichte oder fehlende Gruppen zu erkennen, mögliche Verzerrungen sichtbar zu machen oder nachzuvollziehen, welche Datenpunkte das Modell besonders stark geprägt haben. Solche Analysen helfen einzuschätzen, wie verlässlich und gerecht ein Modell arbeiten kann [17].
4. **LLM-spezifische Erklärbarkeit:** Große Sprachmodelle erzeugen nicht nur Ergebnisse, sondern oft auch begleitende Begründungen. Diese klingen plausibel, sind aber nicht immer

ein verlässlicher Hinweis auf die tatsächlichen inneren Abläufe. Daher befassen sich spezielle Ansätze damit, typische Muster im Modell, etwa die Verarbeitung von Textelementen oder die Gewichtung verschiedener Signale, besser sichtbar und nachvollziehbar zu machen. Ziel ist es, die Funktionsweise solcher Modelle verständlicher zu erklären, ohne sich allein auf ihre selbst formulierten Begründungen zu verlassen [18].

Für XAI existieren mittlerweile ausgereifte Ansätze und Werkzeuge, die sich unabhängig von Branche oder Unternehmensgröße einsetzen lassen [19], [20]. Sie reichen von globalen Modellübersichten, die die generelle Funktionsweise eines Modells sichtbar machen, bis zu lokalen Erklärungsmethoden, die einzelne Entscheidungen im konkreten Einzelfall verständlich machen. Moderne KI-Systeme lassen sich damit nicht nur leistungsfähig, sondern auch überprüfbar, auditierbar und strategisch sicher nutzen. So wird KI von einem schwer einschätzbaren Technologierisiko zu einem verlässlichen Werkzeug.

3.3 Wie kann ein Unternehmen Explainable AI praktisch umsetzen?

Regulatorische Initiativen wie der EU AI Act (siehe Kapitel 2) betonen die Nachvollziehbarkeit von (Hochrisiko-) KI ausdrücklich. Gleichzeitig werden KI-Systeme immer komplexer – dies gilt besonders im Kontext der generativen KI – , was neue Erklärungsmethoden erfordert. Die Forschung arbeitet bereits an noch intelligenteren Erklärungen, von Gegenbeispielen über natursprachliche Erläuterungen bis hin zu automatisierten Audit-Systemen. Für Unternehmen gilt es, frühzeitig eine Strategie für Explainability aufzubauen: Das umfasst die Auswahl geeigneter XAI-Tools, die Schulung von Mitarbeitern im Umgang mit erklärbarer KI und die Verankerung von Erklärbarkeit in den internen KI-Governance-Richtlinien [21]. Zunächst sollte geklärt werden: Soll das Modell *nur* gut funktionieren, oder müssen Entscheidungen auch gegenüber Nutzenden, Aufsichtsbehörden oder Partnern nachvollziehbar gemacht werden? Je nach Zweck variiert der Aufwand. Damit Explainability nicht nur ein abstraktes Prinzip bleibt, sondern im Unternehmen tatsächlich umgesetzt wird, sind einige grundlegende Maßnahmen erforderlich. Diese betreffen sowohl die technische Ebene der Modelle als auch interne Zuständigkeiten und die Kommunikation gegenüber Nutzenden [22]. Die folgenden drei Aufgabenfelder bilden den Kern einer praxisorientierten XAI-Umsetzung.

Technische Erklärbarkeit sicherstellen Unternehmen sollten für jedes KI-System mindestens grundlegende Erklärmethoden implementieren:

- globale Erklärungen (z.,B. Feature-Wichtigkeit),
- lokale Erklärungen (Warum genau diese Entscheidung?),
- Datenanalysen (Bias, Ausreißer, Repräsentativität).

Dokumentation und Verantwortlichkeiten Eine klare interne Dokumentation ist notwendig, damit Verantwortlichkeiten eindeutig zugeordnet werden können und alle Beteiligten nachvollziehen können, wie ein KI-System aufgebaut ist und betrieben wird. Sie bildet die Grundlage für Auditierbarkeit, Qualitätssicherung und regulatorische Anforderungen. Relevante Punkte sind unter anderem:

- *Welche Modelle werden genutzt?*
- *Welche Daten liegen zugrunde?*
- *Welche Erklärmethoden sind verfügbar?*
- *Wer ist verantwortlich für Updates und Monitoring?*

Nutzerorientierte Transparenz Erklärungen zu den Funktionsweisen und Entscheidungen eines KI-Systems müssen so gestaltet sein, dass sie für die jeweiligen Zielgruppen verständlich sind, etwa für Data Analysts, den Kundendienst, das Management oder Endkundinnen und Endkunden. Zu komplexe Erklärungen nützen niemandem.

3.4 Fazit

Explainable AI ist kein Selbstzweck, sondern ein Werkzeug zur Risikominimierung, Qualitätssteigerung und Vertrauensbildung. Für Unternehmen bedeutet XAI vor allem pragmatische Orientierung: einfache Methoden, klare Dokumentation und verständliche Kommunikation. Komplexe Modelle bleiben komplex aber mit den richtigen Erklärmethoden können Unternehmen fundierte Entscheidungen treffen, KI sicher einsetzen und gegenüber internen wie externen Stakeholdern nachvollziehbare Ergebnisse liefern.

Abschließend lässt sich festhalten: Explainable AI ist kein bloßes Add-on, sondern ein Enabler für den erfolgreichen KI-Einsatz in der Praxis. Sie schafft die Vertrauensbasis, auf der Mensch und KI gemeinsam Mehrwert schaffen können. Unternehmen, die ihre KI-Systeme erklärbar gestalten, überzeugen schneller Nutzer und Kunden, sparen Zeit bei Audits und reduzieren Haftungsrisiken. Mit Erklärbarkeit (Explainability) wird aus einer hochentwickelten KI-Lösung letztlich ein akzeptiertes, verlässliches Werkzeug und damit wird ein wesentlicher Schritt hin zu wirklich verantwortungsvoller KI in der Unternehmenswelt gemacht.

4 Green AI

Mit der breiten Einführung großer KI-Modelle rückt die ökologische Dimension von *Responsible AI* zunehmend in den Fokus. Leistungsfähige KI-Modelle gehen häufig mit hohem Rechenaufwand, Energieverbrauch und Kosten einher. Für Unternehmen stellt sich daher die Frage, wie sich der Nutzen von KI mit wirtschaftlichen und ökologischen Zielen in Einklang bringen lässt. Ein zentraler Begriff in diesem Zusammenhang ist *Green AI*: Darunter wird die Forderung verstanden, dass KI-Systeme nicht nur anhand ihrer Leistung und Genauigkeit bewertet werden, sondern auch anhand ihres ökologischen Fußabdrucks. Im Fokus stehen der Energie- und Ressourcenbedarf (und damit verbundene CO_2-Emissionen), insbesondere beim Training und Betrieb sehr großer Modelle. Green AI fokussiert den effizienten Einsatz von Ressourcen: Statt „immer größer und komplexer“ liegt der Schwerpunkt auf Wiederverwendung, Optimierung und maßgeschneiderten Lösungen. Dieses Kapitel zeigt, wie KI-Lösungen so gestaltet werden können, dass sie nicht nur performant, sondern auch nachhaltig und wirtschaftlich tragfähig bleiben.

Schwartz et al. zeigten im Jahr 2019, dass der Rechenaufwand im Deep Learning zwischen 2012 und 2018 um den Faktor 300.000 gestiegen ist [23]. Ebenfalls 2019 quantifizierten Strubell et al. diesen Effekt beispielhaft für NLP-Modelle (Natural Language Processing Modelle). Sie wiesen nach, dass das Training einzelner großer Modelle Emissionen im Bereich mehrerer Tonnen CO_2-Äquivalente verursachen kann [24]. Parallel dazu entwickelt sich zunehmend ein regulatorischer Rahmen: Neben den Erwartungen von Investorinnen und Investoren im Bereich ESG (Environmental, Social und Governance) werden im Kontext des EU AI Acts (siehe Kapitel 2) auch Nachhaltigkeitsaspekte, Transparenzanforderungen und Lebenszyklusbetrachtungen diskutiert [25].

4.1 Green AI in der Unternehmenspraxis

Green AI ist nicht nur eine technische Fragestellung, sondern beeinflusst unmittelbar Kostenstrukturen, Nachhaltigkeitsziele und strategische Entscheidungen im Unternehmen. Der Ressourcenbedarf von KI-Systemen bestimmt, welche Lösungen wirtschaftlich tragfähig sind und wie gut sie sich mit Umwelt- und ESG-Zielen verbinden lassen. In vielen Fällen ist nicht das „größtmögliche“ Modell optimal, sondern eine gezielt optimierte, wiederverwendete oder komprimierte Variante.

Auch falls sich der Nutzen von Green AI nicht immer unmittelbar in Form eines finanziellen Mehrwertes für Unternehmen darstellen lässt, ist er aus langfristiger ökonomischer und ökologischer Sicht strategisch relevant. Unternehmen, die Green-AI-Prinzipien hinreichend berücksichtigen, erfüllen leichter regulatorische Anforderungen, senken mittel- bis langfristig ihre Kosten und stärken gleichzeitig ihre Reputation.

Weitere Effizienzgewinne entstehen, wenn bereits trainierte – insbesondere unternehmensinterne – Modelle gezielt wiederverwendet oder nur inkrementell weitertrainiert werden. Dadurch werden zusätzliche Trainingsläufe vermieden, was Rechenzeit und Energieverbrauch reduziert. Große Modelle lassen sich mit Methoden der Modellkompression verkleinern und auch die Wahl von effizienteren Algorithmen sowie eine auf den Bedarf angepasste Hardware können entscheidende Punkte bei der Senkung des Ressourcenbedarfs sein. Zusätzlich trägt eine systematische Messung des Energieverbrauchs zu mehr Transparenz bei.

Die folgenden Beispiele zeigen typische Situationen, in denen Green-AI-Prinzipien eine besondere Rolle spielen und in denen sich die in den folgenden Abschnitten beschriebenen Ansätze zur Wieder- und Weiterverwendung von Modellen, zur Modellkompression und zu sonstigen Maßnahmen konkret anwenden lassen:

- Einführung eines großen Sprachmodells für interne Wissenssuche, bei der statt eines von Grund auf neu trainierten Modells ein bestehendes Sprachmodell per Transfer Learning und LoRA-Fine-Tuning an Unternehmensdaten angepasst wird. Durch gezielte Modellkompression und eine geeignete Betriebsumgebung werden sowohl Infrastrukturkosten als auch Energieverbrauch reduziert.
- Einsatz von KI für vorausschauende Wartung (Predictive Maintenance), bei dem ein aus einem größeren Basismodell abgeleitetes und komprimiertes Modell direkt auf Edge-Geräten oder Industrie-PCs ausgeführt wird, sodass Rechenaufwand, Bandbreite und Energieverbrauch im laufenden Betrieb gering bleiben und gleichzeitig Wartungszyklen optimiert werden können.
- Aufbau einer internen Modellsuite, in der ein zentrales Basismodell für mehrere Produktlinien wiederverwendet wird und pro Anwendungsfall nur klein dimensionierte Adapter oder feinjustierte Teilmodelle trainiert werden, anstatt für jede Aufgabe ein vollständig neues Modell aufzubauen.
- Nutzung von Cloud-Ressourcen für das (Weiter-)Training von Modellen, bei der durch optimierte Trainingsalgorithmen, passende Hardwareprofile und die gezielte Begrenzung von Trainingsläufen sowohl Kosten als auch CO_2-Bilanz aktiv gesteuert und mithilfe von Monitoring-Werkzeugen messbar gemacht werden.
- Berücksichtigung von KI-spezifischen Verbräuchen in ESG-Reporting und Nachhaltigkeitsberichten, indem Energieverbrauch und Emissionen zentraler Modelle systematisch erfasst und die Effekte von Modellrecycling, Kompression und weiteren Effizienzmaßnahmen transparent ausgewiesen werden.

4.2 Wieder- und Weiterverwendung von Modellen

Ein besonders wirkungsvoller Ansatz im Unternehmensumfeld ist die Nutzung bereits vortrainierter Modelle anstatt Modelle von Grund auf neu zu trainieren. Dies reduziert Entwicklungszeiten, senkt

Infrastrukturkosten und verringert den Energiebedarf erheblich. Studien aus verschiedenen Domänen zeigen, dass Transfer Learning mit vergleichsweise wenigen zusätzlichen Trainingsschritten zu hoher Performanz führt, während der Energiebedarf im Vergleich zum Neutraining drastisch sinkt [26], [27]. Praxisleitfäden, etwa aus dem *Green Software Patterns Catalog*, empfehlen explizit, vortrainierte Modelle als Standardoption zu prüfen, um den Carbon Footprint von KI-Projekten zu senken [28]. Dabei treten mitunter neue Zielkonflikte auf: Die Zentralisierung auf wenige Basis- oder Foundation-Modelle kann Abhängigkeiten von Plattformanbietern verstärken und wirft Fragen zu Vendor Lock-in, Datensouveränität und Transparenz der zugrunde liegenden Modelle auf. Grundsätzlich ist auch zu unterscheiden, ob ein Modell in einem ähnlichen Kontext eingesetzt oder sein Kontext erweitert werden soll.

4.2.1 Transfer Learning

Die Übertragung von vorhandenem Kontextwissen auf einen ähnlichen Anwendungsbereich wird als *Transfer Learning* bezeichnet. Dies ist vor allem bei kleinen Datensätzen von Vorteil, wenn die Datenmenge für ein vollständiges Training nicht ausreicht. Zunächst wird das Modell mit einem größeren, ähnlich gelagerten Datensatz trainiert (z. B. deutsche Standardsprache), anschließend erfolgt ein Fine-Tuning mit domänenspezifischen Daten (z. B. Fachsprache). Eine moderne und effiziente Technik dafür ist *Low-Rank Adaptation (LoRA)*. Hierbei wird im Gegensatz zum klassischen Fine-Tuning nur ein Teil der Modellparameter angepasst, meist einzelne Schichten, was ein schnelles und ressourcenschonendes Fine-Tuning großer Modelle ermöglicht. LoRA basiert auf der Beobachtung, dass Gewichtsmatrizen während des Trainings eine niedrige *„intrinsische" Dimension* einnehmen, was eine einfache Niedrigrangzerlegung erlaubt. Studien zeigen, dass LoRA die Anzahl der zu trainierenden Parameter für ein GPT-3-Modell um den Faktor 10.000 reduziert und den GPU-Speicherbedarf um das Dreifache verringert [29]. Für Unternehmen bedeutet dies, dass bestehende Modelle sowohl aus Kosten- als auch Umweltgesichtspunkten sinnvoll wiederverwendet und angepasst werden können. Voraussetzung ist eine hinreichend ähnliche Datenbasis, um negativen Transfer zu vermeiden. Negativer Transfer meint hier den Leistungseinbruch, der entsteht, wenn ein vortrainiertes Modell auf neuen Daten genutzt wird, die zu verschieden zu den ursprünglichen Trainingsdaten sind. In diesem Fall kann es nötig sein, ein Modell gänzlich neu anzulernen.

4.2.2 Continuous Learning

Erfolgt eine Erweiterung des Modellkontexts um neue Anwendungsfälle, wird dies als inkrementelles Lernen oder *Continuous Learning (CL)* bezeichnet (vgl. [30]). Idealerweise werden die Anforderungen frühzeitig definiert, sodass geeignete CL-Methoden direkt von Beginn an eingesetzt werden können. Im Kontext von Sprachmodellen bezeichnet CL die Fähigkeit eines Modells, beständig Wissen zu erweitern, indem neues Wissen gezielt mit bereits bestehendem Wissen verknüpft wird, sodass das Modell sich kontinuierlich verbessert und an neue Sprache und Inhalte anpasst, ohne zuvor Erlerntes

zu vergessen. Eine große Herausforderung hierbei ist das sogenannte *Catastrophic Forgetting (CF)*, bei dem frühere Lerninhalte überschrieben werden, sobald neue Daten integriert werden [31]. Dies führt zu einem bekannten Dilemma [32]: Einerseits sollen neue Tasks hinreichend flexibel angelernt werden können, d.h. das Modell muss hinreichend plastisch sein. Andererseits soll aber auch bereits vorhandenes Wissen hinreichend konserviert werden, wodurch das Modell auch eine gewisse Stabilität gewährleisten muss. Da sich beide Aspekte gegenseitig beeinflussen, muss hier eine geeignete Balance gefunden werden. Dies erfolgt über Regularisierungsmethoden, Replay-Verfahren und architekturbezogene Anpassungen [30]. Wegen des hohen Ressourcenbedarfs für fortlaufendes Training großer Sprachmodelle ist eine effiziente Ressourcennutzung essenziell, was für kleinere Organisationen eine Herausforderung darstellen kann. Kleinere, spezialisierte oder lokal gehostete Modelle können hier eine Alternative darstellen, ebenso wie gezielte Modellkompression.

4.3 Modellkompression

Unter Modellkompression werden Techniken zusammengefasst, die gezielt zur Verkleinerung großer, ressourcenintensiver KI-Modelle angewendet werden, um deren Speicherbedarf und Rechenanforderungen zu reduzieren sowie gleichzeitig die Leistungsfähigkeit möglichst gut zu erhalten [33]. Ziel ist es, KI-Modelle auch in ressourcenlimitierten Umgebungen wie mobilen Geräten oder Edge-Computing effizient nutzbar zu machen, wodurch es sich auch um eine wesentliche Schlüsseltechnik zur Demokratisierung von KI handelt. Wichtige Kompressionsverfahren umfassen unter anderem:

- **Pruning:** Entfernen unnötiger Verbindungen oder ganzer Schichten eines Netzes [34],[35],[33]. Praktisch eignet sich Pruning insbesondere, um Modelle für Edge- und Predictive-Maintenance-Szenarien zu verschlanken, sodass Zustandsüberwachung direkt auf Industrie-PCs oder eingebetteten Systemen mit geringerer Latenz und weniger Energieverbrauch möglich wird.
- **Quantisierung:** Reduktion der numerischen Präzision der Modellparameter (z. B. von 32-Bit auf 8-Bit) [36],[33]. Quantisierung hilft vor allem dort, wo vorhandene Hardware weitergenutzt werden soll, etwa bei internen Sprachmodellen zur Wissenssuche: Durch reduzierte numerische Präzision sinken Speicher- und Rechenbedarf, ohne dass zwingend ein spürbarer Qualitätsverlust entsteht.
- **Knowledge Distillation:** Übertragung von Wissen eines großen Modells auf ein kleineres [37],[38],[33]. Knowledge Distillation kann genutzt werden, um aus einem großen, universellen Basismodell mehrere kleinere, auf konkrete Aufgaben zugeschnittene Modelle abzuleiten, die sich effizient in einer internen Modellsuite betreiben und gezielt für verschiedene Fachbereiche oder Produkte einsetzen lassen.
- **Low-Rank-Dekomposition:** Zerlegung komplexer Gewichtsmatrizen in kleinere Komponenten (vgl. LoRA unter 4.2) [29],[33]. Low-Rank-Dekomposition und darauf aufbauende Verfahren wie LoRA sind insbesondere dann hilfreich, wenn ein gemeinsames Grundmodell für viele

Anwendungsfälle erhalten bleiben soll, Anpassungen aber in Form schlanker Adapter erfolgen sollen, die sich schnell trainieren und ressourcenschonend ausrollen lassen.

In der Unternehmenspraxis entfalten diese Modellkompressionstechniken ihre Nützlichkeit vor allem dadurch, dass sie leistungsfähige Modelle unter realen Randbedingungen handhabbar machen. Durch das gezielte Verkleinern von Modellen können Speicherbedarf, Rechenaufwand und Energieverbrauch deutlich reduziert werden, ohne die Modellgüte unverhältnismäßig stark zu beeinträchtigen. Dies erleichtert nicht nur den Betrieb auf bestehender Infrastruktur oder ressourcenbeschränkten Geräten (z. B. Edge- oder Industrie-Hardware), sondern senkt auch laufende Kosten und unterstützt Nachhaltigkeits- und ESG-Ziele.

Gleichzeitig ermöglichen komprimierte Modelle, dass ein einzelnes großes Basismodell in Form kleinerer, spezialisierter Varianten für unterschiedliche Anwendungsfälle genutzt werden kann. Statt für jede Aufgabe ein neues, großes Modell zu trainieren, lassen sich so mehrere schlanke Modelle betreiben, die auf einem gemeinsamen Wissensstand aufbauen und effizient angepasst werden können. Modellkompression ist damit ein zentraler Baustein, um Wieder- und Weiterverwendung von Modellen, wirtschaftliche Tragfähigkeit und ökologische Verantwortung zusammenzubringen.

4.4 Weitere Maßnahmen

Neben der Weiter- und Wiederverwendung von Modellen sowie der gezielten Verkleinerung großer Modelle durch Modellkompression gibt es weitere Maßnahmen zur effizienteren Nutzung von Ressourcen. Dazu zählen vor allem die Entwicklung und Anwendung effizienterer Algorithmen, die den Rechenaufwand und somit auch den Ressourcenbedarf deutlich reduzieren können. Solche Algorithmen umfassen beispielsweise adaptive und selbstlernende Systeme, die sich dynamisch an veränderte Bedingungen anpassen und somit Ressourcen optimal einsetzen können. Ergänzend spielen Hardware-Optimierungen eine wesentliche Rolle, beispielsweise spezialisierte Chips, die explizit für bestimmte KI-Aufgaben entwickelt wurden und somit besonders energieeffizient arbeiten. Ein weiterer Punkt ist die Messung und Überwachung des Energieverbrauchs von KI-Systemen, wodurch gezielt Einsparpotenziale erkannt werden können. Alle in diesem Abschnitt genannten Maßnahmen können zusätzlich zur nachhaltigen Gestaltung von KI-Anwendungen beitragen und auch kombiniert werden.

4.5 Fazit

Die Integration von Responsible AI und insbesondere Green AI ist für Unternehmen heute unerlässlich. Aktuelle Diskussionen um stark steigende Emissionen großer Cloud-Anbieter – etwa der deutliche Anstieg bei Google aufgrund wachsender KI-Last – verdeutlichen, dass Green AI kein Nischenthema ist, sondern ein zentraler Bestandteil verantwortungsvoller Unternehmensstrategien [39]. Nachhaltigkeitsstrategien fördern langfristig Wettbewerbsfähigkeit und stärken das Unternehmensimage.

Empfohlene Ansätze aus aktuellen Leitfäden zielen auf die Reduktion von Umweltbelastungen wie CO_2-Ausstoß und Ressourcenverbrauch ab und ermöglichen zugleich wirtschaftliche Vorteile. Anstatt Modelle neu zu entwickeln, sollte immer vorab geprüft werden, ob nicht bereits vorhandene Modelle gezielt recycelt und durch Fine-Tuning, spezielle Lernmethoden, dynamische Architektur und Maßnahmen zur Erweiterung bzw. Komprimierung von Modellen effizient und vorausschauend angepasst werden können. Optimierte Algorithmen und passende Hardware tragen zusätzlich zur Energieeinsparung bei. Verantwortungsvolles Monitoring schafft dabei eine transparente Grundlage für die Bewertung unternehmerischer und gesellschaftlicher Kosten und stärkt das Vertrauen aller Beteiligten. Letztlich ist die Balance zwischen Leistungsfähigkeit und Ressourceneinsatz entscheidend: Während große Modelle mehr Ressourcen benötigen, sind kleinere, modulare und komprimierte Ansätze häufig ausreichend. Eine starke Zentralisierung auf wenige Basis-Modelle wirft zudem Fragen zu Abhängigkeiten, Datensouveränität und Transparenz auf. Für eine erfolgreiche Integration von Green AI sollten Unternehmen Energieverbrauch und Emissionen gezielt nachhalten, interne Richtlinien für Modell-Recycling etablieren sowie Entscheidungen immer mit vorausschauendem Blick auf Genauigkeit, Kosten, Umweltwirkung und Risiken treffen. Ansätze wie LoRA bieten Wege für eine modulare, flexible und ressourcenschonende Modellentwicklung. Green AI bewirkt somit nicht nur ökologische Verbesserungen, sondern ermöglicht auch Kosteneinsparungen, stärkt die Wettbewerbsfähigkeit und sichert die Zukunftsfähigkeit von Unternehmen.

5 Lokale Modelle (On-Prem/Edge) im Responsible-AI-Kontext

Vor der breiten Anwendung großer Cloud-basierter Sprachmodelle war der Einsatz von KI in Unternehmen in vielen Fällen bereits ein vollständig lokales Thema z.B. in Form von eingebetteten Systemen in Maschinen, Fahrzeugen oder Produktionsanlagen. Mit der Verfügbarkeit leistungsfähiger generativer Modelle hat sich der Fokus jedoch verschoben: Viele Organisationen testen zunächst Cloud-LLMs, stehen dann aber vor der strategischen Frage, welche Teile ihrer KI-Landschaft langfristig in der Cloud, on-premise oder direkt an der „Edge“ betrieben werden sollen. Gerade für kleine und mittlere Unternehmen (KMU) geht es dabei nicht nur um technische Leistungsfähigkeit, sondern auch um Fragen der Datensouveränität, der Abhängigkeit von Anbietern und der langfristigen Betriebskosten.

Unter „lokalen Modellen“ verstehen wir in diesem Kapitel KI-Systeme, die vollständig in der eigenen Infrastruktur (On-Premise) oder direkt auf Geräten und Maschinen (Edge) ausgeführt werden, im Gegensatz zu rein Cloud-basierten KI-Diensten. Lokale LLMs und Edge-AI ermöglichen es, sensible Daten im eigenen Verantwortungsbereich zu halten, Entscheidungen mit minimaler Latenz zu treffen und Modelle tiefgreifend an unternehmensspezifische Anforderungen anzupassen. Gleichzeitig bringen sie eigene Herausforderungen mit sich: Investitionen in Hardware, Betriebs- und Wartungsaufwand sowie die Notwendigkeit, internes Know-how für Auswahl, Betrieb und Absicherung dieser Systeme aufzubauen.

Im Responsible-AI-Kontext sind lokale Modelle ein wichtiger Baustein, um mehrere der zuvor diskutierten Dimensionen zusammenzuführen: Sie unterstützen Datensouveränität und Datenschutz (z. B. im Lichte von DSGVO und EU AI Act), ermöglichen domänenspezifische Optimierungen und können, in Kombination mit geeigneten Modellgrößen und -architekturen, einen Beitrag zu Green AI leisten, indem Rechenaufwand und unnötige Datenbewegungen reduziert werden. Gleichzeitig müssen Unternehmen abwägen, wann lokale Lösungen tatsächlich Vorteile bringen und in welchen Szenarien Cloud-Angebote weiterhin sinnvoll sind.

5.1 Definitionen lokaler Modelle

Lokale KI-Modelle sind KI-Systeme, die vollständig in der eigenen Infrastruktur eines Unternehmens oder direkt auf Endgeräten (Edge-Geräten) betrieben werden – im Gegensatz zu Cloud-basierten KI-Diensten externer Anbieter. Bei einem **On-Premise-Modell** läuft das KI-Modell auf eigenen Servern oder im Rechenzentrum der Organisation, wodurch alle Modelldaten (Gewichte, Eingaben und Ausgaben) im Unternehmensnetz verbleiben [40]. Dies gibt dem Unternehmen volle Kontrolle über das Modell inklusive Updates, Konfiguration und Skalierung und garantiert, dass sensible Daten das eigene Netzwerk nicht verlassen. Dadurch entfallen Abhängigkeiten von externen KI-Services

wie Vendor-Lock-in, Ausfällen oder plötzliche API-Preisänderungen von Drittanbietern [41]. **Edge-Modelle** hingegen sind direkt auf dezentralen Geräten oder Maschinen z. B. in Fertigungsanlagen, Fahrzeugen oder IoT-Geräten integriert [42]. Sie führen KI-Berechnungen vor Ort in Echtzeit durch, ohne kontinuierliche Verbindung zur Cloud. Dies reduziert Latenzzeiten, spart Bandbreite und erhöht die Ausfallsicherheit, da die Inferenz auch bei begrenzter oder unterbrochener Internetanbindung weiterläuft.

Im Vergleich zu Cloud-basierten KI-Diensten (bei denen ein externer Anbieter das Modell in seiner Cloud bereitstellt) bieten lokale Modelle somit insbesondere Vorteile bei Datenschutz, Kontrolle und Anpassbarkeit [43]. Cloud-LLMs zeichnen sich dagegen durch schnelle Einsatzbereitschaft und einfache Skalierbarkeit aus. Neue KI-Funktionen lassen sich dort sofort per API einbinden, ohne eigene Hardware beschaffen zu müssen. Allerdings müssen Unternehmen bei Cloud-Lösungen in Kauf nehmen, dass Daten das eigene sichere Netz verlassen und auf fremden Servern verarbeitet werden. Die folgende Gegenüberstellung verdeutlicht die Unterschiede:

- **Datenkontrolle & Sicherheit**: Lokale LLMs bieten volle Datensouveränität – alle Datenverarbeitungen erfolgen im eigenen, kontrollierten Umfeld, was sie ideal macht für sensible oder regulierte Informationen. Bei Cloud-LLMs verbleibt die Sicherheitsverantwortung größtenteils beim Anbieter; trotz hoher Cloud-Sicherheitsstandards werden die Daten über das Internet an Rechenzentren übertragen und unterliegen dortigen Zugriffsrisiken
- **Anpassbarkeit & Integration**: On-Premise-Modelle lassen sich tiefgreifend anpassen, z.B. durch Fine-Tuning mit firmeneigenen Daten, eigene Tokenizer oder Modifikationen an der Modellarchitektur. Diese Flexibilität erlaubt es, Domänenwissen und firmenspezifische Anforderungen direkt im Modell abzubilden. Cloud-Angebote bieten hingegen meist nur eingeschränkte Anpassungsmöglichkeiten (oft beschränkt auf vom Anbieter bereitgestellte Parameter oder Fine-Tuning-APIs).
- **Leistung & Latenz**: Lokal betriebene KI-Modelle reagieren oft mit minimaler Latenz, da keine Netzwerkübertragung nötig ist. Besonders bei Echtzeitanwendungen (etwa in Fahrzeugen oder Produktionsanlagen) ist dies kritisch. Cloud-Modelle weisen demgegenüber eine netzwerkbedingte Verzögerung auf und ihre Performance hängt von der Internetanbindung sowie der Auslastung des Anbieter-Servers ab.
- **Skalierbarkeit & Betriebsaufwand**: Cloud-Services erlauben eine sofortige Skalierung nach Bedarf (automatisches Hinzufügen von Rechenressourcen) und erfordern anfangs geringere Investitionen. On-Premise-Lösungen bieten zwar konsistente, dedizierte Rechenleistung, müssen aber bei wachsendem Bedarf manuell durch zusätzliche Hardware ausgebaut werden. Das bedeutet höhere Initialkosten (für Server, GPUs, Speicher) und laufenden Wartungsaufwand im eigenen Haus. Dafür können bei dauerhaft intensiver Nutzung die Gesamtkosten planbarer und langfristig sogar geringer sein als bei nutzungsbasierten Cloud-Gebühren

5.2 Datensouveränität, Datenschutz und Compliance

Ein entscheidender Vorteil lokaler KI-Modelle ist die Wahrung der Datensouveränität. Unternehmen behalten die volle Kontrolle darüber, wo und wie ihre Daten durch KI verarbeitet werden. Im Gegensatz zu Cloud-Diensten, bei denen Eingabedaten das Unternehmen verlassen und auf fremder Infrastruktur verarbeitet werden, bleiben bei einem lokal betriebenen Modell sämtliche Daten im eigenen Verantwortungsbereich. Dies erleichtert die Umsetzung von Privacy-by-Design: Datenschutz wird von vornherein technisch integriert. So können selbst streng vertrauliche Informationen wie geheime Entwicklungsdokumente in einer Firma oder Patientendaten im Krankenhaus durch KI-Systeme analysiert werden, ohne dass ein Risiko des externen Abflusses besteht [44]. Konkret bedeutet das: Wenn KI-Systeme auf eigenen Servern oder in einem inländischen Rechenzentrum laufen, können Firmen sicherstellen, dass alle personenbezogenen Daten gemäß DSGVO geschützt bleiben. Rechtliche Grauzonen beim Einsatz von Cloud-KI (etwa ob Anbieter im Drittland Zugriff haben könnten) werden so vermieden. Die **rechtliche Compliance** in Bezug auf Datenschutz ist dadurch deutlich einfacher zu erreichen – insbesondere in datensensiblen Branchen (Gesundheitswesen, Finanzen, öffentlicher Sektor), wo Cloud-Lösungen oft an regulatorische Grenzen stoßen.

5.3 Domänenspezifische Optimierung: Fine-Tuning und RAG

Neben dem Datenschutzaspekt erlaubt der lokale Betrieb auch, domänenspezifisches Wissen optimal zu integrieren. Unternehmen können ihre KI-Modelle gezielt auf den eigenen Anwendungsfall zuschneiden. Während dies bei klassischen KI-Modellen im Vergleich zu generativer KI häufig der Normalfall ist, werden generische Cloud-LLMs darauf trainiert sind, breit anwendbar zu sein. Lokale LLMs können stattdessen mit branchenspezifischen Daten, Fachterminologie und Dokumenten angereichert werden. Auch für Edge-KI Anwendungen kann es ebenfalls allgemeinere oder bereits trainierte Modelle geben, die jedoch in den meisten Fällen nicht direkt verwendet werden können und an die spezifischen Daten z.B. von Sensoren oder den zu klassifizierenden Objekte angepasst werden müssen. Dadurch befindet man sich eher im Bereich der Wieder- und Weiterverwendung von Modellen (s. Kapitel 4.2), weshalb der Fokus dieses Abschnitts auf generativer KI und LLMs liegt. Für generative KI gibt es zwei Hauptansätze für domänenspezifische Optimierung: **Fine-Tuning** des Modells mit eigenen Trainingsdaten und **Retrieval-Augmented Generation (RAG)** mittels unternehmenseigener Wissensdatenbanken [45].

Durch **Fine-Tuning** auf firmeninterne Textcorpora lässt sich ein Sprachmodell beispielsweise auf juristische Fachsprache, medizinische Befunde oder technischen Jargon „einstimmen“. Die Folge ist, dass das Modell präzisere und relevantere Ergebnisse in diesem Spezialgebiet liefert als ein allgemeines Modell. Zielgerichtetes Fine-Tuning vortrainierter Sprachmodelle können massive Performance-Gewinne bei spezifischen Textklassifizierungsaufgaben bringt. Dieses Prinzip findet heute bei LLMs breite Anwendung. On-Premise-Modelle vereinfachen solche Anpassungen erheblich:

Da die Modellgewichte verfügbar sind, kann man beliebige zusätzliche Trainingsdaten einspielen und sogar die Architektur anpassen oder spezielle Tokenizer verwenden, um firmenspezifische Begriffe optimal zu verarbeiten. Ein Cloud-LLM bietet diese Tiefe der Anpassung oft nicht; dort ist man auf die vordefinierten Funktionen des Anbieters beschränkt. In einer lokalen Umgebung hingegen kann das KI-Team das Modell beispielsweise mit historischen Unternehmensdaten weitertrainieren, sodass es firmeneigene Produktnamen, interne Abkürzungen und Richtlinien „versteht“. Auch mehrsprachige Anpassungen sind möglich: Hat eine Organisation z.B. viel deutschsprachiges Wissen, kann sie ein mehrsprachiges Open-Source-Modell auf Deutsch feinjustieren oder ein bereits deutsch optimiertes Modell einsetzen, um bessere Ergebnisse in unserer Sprache zu erzielen. Die europäische KI-Startup-Szene bringt solche Modelle hervor – etwa Mistral AI aus Frankreich, deren LLM-Variante gezielt auf europäische Mehrsprachigkeit und Datenschutz ausgelegt ist. Deren Modelle laufen auf europäischer Infrastruktur und enthalten beim Training umfangreiche deutsch-, französisch- usw. sprachige Daten, was zu kulturell und fachlich passenderen Ergebnissen führt.

Eine weitere Stärke von On-Premise-Modellen ist die Integration ins bestehende IT-Ökosystem, wodurch Unternehmen ihre LLMs mit internen Datenquellen wie Datenbanken, Wissensmanagement-Systeme oder Dokumentenarchive koppeln können. Mit **Retrieval-Augmented Generation (RAG)** kann etwa ein lokales Modell bei Bedarf die firmeneigene Wissensdatenbank abfragen, um aktuelle Fakten oder Richtlinien in seine Antwort einfließen zu lassen. Auf diese Weise wird das LLM gewissermaßen zu einem unternehmensspezifischen Wissensspeicher, der die firmeneigene Terminologie und Praxis beherrscht. All das geschieht außerdem innerhalb der geschützten Umgebung der Organisation, sodass sämtliche Daten intern auf den eigenen Servern bleiben und dadurch die eigene Datensouveränität sicherstellt. Der Vorteil von RAG im Vergleich zum Fine-Tuning besteht also darin, dass sich der Datenbestand kontinuierlich erweitern lässt, während ein Fine-Tuning auf einem bestimmten, zum Trainingszeitpunkt eingefrorenen Wissensstand basiert. Wenn sich das verfügbare Wissen eines Unternehmens erweitert, genügt mit RAG eine Aktualisierung der Datenbanken und es muss nicht das gesamte Modell erneut trainiert werden. Diese Freiheit führt im Umkehrschluss aber auch dazu, dass man mit RAG im Vergleich zu einem Fine-Tuning weniger Einfluss auf die „natürlichen“ bzw. antrainierten Fähigkeiten des Modells hat. Man fügt jedem Prompt lediglich einen domänenspezifischen Kontext hinzu, wodurch sich zusätzlich auch die notwendige Bandbreite im Unternehmensnetzwerk erhöhen kann. Auf Grund der hohen Flexibilität lässt sich RAG aber auch technisch einfach bestehende On-Premise Architekturen integrieren, sodass Pilotprojekte mit RAG schneller ausgerollt werden können.

Die beiden Ansätze lassen sich natürlicherweise auch **kombinieren**, indem konstantes bzw. wenig veränderliches und essentielles Wissen sowie striktere Vorgaben an Stil und Sprache für ein Fine-Tuning verwendet werden und veränderliches bzw. erweiterbares Wissen mittels RAG aus Wissensdatenbanken abgefragt wird. Zusammengefasst stärken lokale Modelle also die Datenhoheit und ermöglichen es, KI maßgeschneidert auf Branchen- und Unternehmensanforderungen

zuzuschneiden. Durch Privacy-by-Design können vertrauliche Informationen sicher genutzt werden, und dank Feinanpassung an Domänenwissen erzielen diese Modelle oft höhere Genauigkeit und Relevanz in speziellen Aufgaben als generische KI-Dienste. Dies ist ein wesentlicher Faktor, um KI im Unternehmenskontext verantwortungsvoll und effektiv einzusetzen.

5.4 Einsatzbereiche lokaler Modelle

Trotz des höheren Betriebs- und Wartungsaufwands entscheiden sich viele Unternehmen für lokale KI, wenn **Datenschutz/Datensouveränität**, **geringe Latenz** oder **domänenspezifische Anpassungen** im Vordergrund stehen. Die folgenden Beispiele skizzieren typische Einsatzfelder für lokale LLMs (On-Premise) und Edge-AI in der Unternehmenspraxis:

- **Industrie 4.0:** In Fertigungs- und Produktionsumgebungen werden Edge-KI-Modelle insbesondere für *Qualitätskontrolle* (z. B. visuelle Inspektion mit Kameras direkt an der Linie) und *prädiktive Wartung* (engl. predictive maintenance) eingesetzt. Sensordaten und Bilddaten werden vor Ort ausgewertet, um frühzeitig Verschleiß oder Anomalien zu erkennen, ohne große Datenmengen kontinuierlich in die Cloud übertragen zu müssen. Das ermöglicht Echtzeitreaktionen, reduziert Bandbreitenbedarf und erhöht die Robustheit bei Verbindungsstörungen.
- **Automotive:** Moderne Fahrzeuge integrieren KI-Systeme lokal an Bord, etwa für Fahrerassistenzsysteme und Funktionen mit (teil-)autonomem Verhalten. Hier ist Edge-AI zentral, weil sicherheitskritische Entscheidungen in Millisekunden getroffen werden müssen (z. B. Objekterkennung, Spurhalten, Notbremsung) und Cloud-Latenzen nicht tolerierbar sind. Die lokale Verarbeitung von Kamera- und Sensordaten erhöht zudem die Ausfallsicherheit (z. B. in Funklöchern) und schützt die Privatsphäre, da Rohdaten der Fahrumgebung das Fahrzeug nicht verlassen.
- **Kundenservice & interne Assistenzsysteme:** Viele Organisationen setzen auf sprachbasierte Assistenzsysteme, etwa Chatbots für Kundenanfragen oder interne Wissensassistenten. Beim lokalen Hosting von LLM-basierten Systemen können auch vertrauliche Inhalte (z. B. Kundendaten, interne Dokumente, geschützte Prozessinformationen) verarbeitet werden, ohne Daten an Dritte weiterzugeben. Das ist besonders relevant in regulierten Bereichen (z. B. Finanzsektor, öffentlicher Sektor), in denen Datenflüsse und Zugriffsrechte streng kontrolliert werden müssen. Typische Einsatzfälle sind die Unterstützung im Technical Support, die automatische Beantwortung wiederkehrender Anfragen oder die Recherche in internen Wissensbeständen.
- **Gesundheitswesen:** In Kliniken und Telemedizin unterstützen lokal betriebene KI-Modelle u. a. medizinische Assistenzsysteme (z. B. Entscheidungsunterstützung bei Diagnostik/Therapie) sowie die Patientenüberwachung über vernetzte Medizingeräte. Edge-AI kann Vitaldaten in Echtzeit auswerten und bei Auffälligkeiten unmittelbar alarmieren. Durch die Verarbeitung

innerhalb der Krankenhausinfrastruktur lassen sich Datenschutz- und Schweigepflichtanforderungen besser umsetzen (Privacy-by-Design), sodass KI-Anwendungen auch in besonders sensiblen Kontexten compliance-gerecht betrieben werden können.

In der Praxis ist dabei häufig ein **hybrider Ansatz** sinnvoll: Für schnelle Pilotierung und erste Experimente kann ein Cloud-Modell Vorteile bieten (Skalierung, geringe Einstiegshürden). Für produktive, sensible oder latenzkritische Use Cases werden anschließend häufig kleinere, feinjustierte Modelle on-premise oder an der Edge betrieben. Damit lassen sich Innovationsgeschwindigkeit und Kontrolle über Daten, Betrieb und Compliance in einer Gesamtarchitektur zusammenführen.

5.5 Beitrag zu Green-AI: Energieeffizienz und Nachhaltigkeit

Wie bereits in Kapitel 4 dargestellt, sind nicht nur die Rechenleistung sondern auch die „Klima-Bilanz“ wichtig beim Einsatz von KI-Modellen. Edge-AI-Ansätze und kleine, spezialisierte On-Premise LLM/LMM sind hierbei also wichtiger Bausteine für mehr Nachhaltigkeit. Anstatt für jede Aufgabe ein riesiges Modell von Grund auf neu zu trainieren, kann die Wiederverwendung vortrainierter Modelle sowie deren gezielte Feinanpassung den Rechenaufwand drastisch senken. Es empfiehlt sich daher vortrainierte KI-Modelle als Standardoption in Projekten zu prüfen, um den Carbon Footprint zu reduzieren. Für Unternehmen ergibt sich so ein doppelter Nutzen: Durch die Nutzung vorhandener Basismodelle (z.B. unternehmensweite Foundation-Modelle für Text-, Bild- oder Prognoseaufgaben) lassen sich Infrastrukturkosten einsparen, die Time-to-Market verkürzen und gleichzeitig Emissionen reduzieren. Die zentrale Nutzung einiger weniger großer Modelle birgt zwar Risiken (etwa Abhängigkeit von einzelnen Anbietern und mögliche Transparenzprobleme bei deren Training), doch dem kann man mit distillierten oder spezialisierten kleineren Modellen begegnen. Diese sind auf spezielle Aufgaben oder Domänen optimiert und können oft mit einem Bruchteil der Rechenressourcen ähnlich gute Ergebnisse erzielen. Ein bewusster Balanceakt ist nötig, um für jede Anwendung die kleinstmögliche Modellgröße zu wählen, die noch ausreichende Leistung bringt. Damit wird unnötiger Rechenaufwand vermieden.

Zusätzlich entfällt bei **Edge-AI** die permanente Datenübertragung an zentrale Server: Werden Daten direkt an der Quelle verarbeitet (z. B. Sensordaten in einer Anlage oder Videodaten in einem Fahrzeug), spart dies auch die Energie, die für das Senden, Empfangen und Speichern großer Datenströme in Rechenzentren anfällt. In der Automobilindustrie etwa reduzieren lokale KI-Systeme den Bedarf, Fahrzeugdaten in die Cloud zu schicken, und **senken so Kosten und Energie für Datenübertragung** und schnelle Diagnosen können direkt vor Ort stattfinden. Ferner erlauben Edge-Geräte oft eine **optimierte Hardware-Beschleunigung**: Spezialisierte KI-Chips (z. B. NPUs in Smartphones oder FPGAs/ASICs in IoT-Geräten) können Inferenz mit deutlich geringerem Stromverbrauch pro Operation durchführen als allgemeine Rechenzentren-CPUs oder

GPUs. Dadurch werden KI-Funktionen energieeffizient „an der Edge“ ausgeführt, was in Summe den Stromverbrauch senken kann, insbesondere wenn viele Geräte parallel arbeiten.

Zusammenfassend lässt sich sagen, dass lokale und kleinere KI-Modelle die Nachhaltigkeit durch **geringeren Ressourcenverbrauch**, Vermeidung unnötiger Datenbewegungen und effizientere Hardware-Nutzung fördern. Sie ermöglichen es Unternehmen, KI-Innovationen umzusetzen und gleichzeitig den CO_2-Fußabdruck ihrer KI-Systeme zu minimieren – ein wichtiges Kriterium verantwortungsvoller Unternehmensführung.

6 Zusammenfassung & Ausblick

Responsible AI beschreibt den Anspruch, KI-Systeme so zu entwickeln und zu nutzen, dass sie *rechtssicher*, *nachvollziehbar*, *nachhaltig* und *datensouverän* in die Wertschöpfung integriert werden können. Im Unternehmensalltag heißt das: KI ist nicht nur ein IT-Thema, sondern beeinflusst Prozesse, Produkte, Verantwortung und Reputation gleichermaßen. Der entscheidende Perspektivwechsel lautet daher nicht mehr „Setzen wir KI ein?“, sondern „Wie setzen wir KI verantwortungsvoll, regelkonform und dauerhaft tragfähig ein?“. Responsible AI bündelt rechtliche Vorgaben, ethische Leitplanken, ökologische Ziele und gesellschaftliche Erwartungen zu einem handhabbaren Rahmen, der sowohl Risiken reduziert als auch Wertschöpfung ermöglicht: Wer systematisch vorgeht, senkt Haftungs- und Reputationsrisiken, verbessert die Entscheidungsqualität und schafft die Voraussetzungen, KI verlässlich zu skalieren.

Im europäischen Kontext ist der EU AI Act dabei ein zentraler Orientierungspunkt, weil er Pflichten risikobasiert an der Kritikalität einer Anwendung ausrichtet. Für Unternehmen bedeutet das vor allem, KI-Anwendungen sauber einzuordnen: Welche Einsätze sind unzulässig, welche potenziell hochriskant und welche fallen in weniger kritische Bereiche? Eng damit verbunden ist die Rollenklärung, ob ein Unternehmen als *Anbieter* (entwickelt, modifiziert oder bringt KI unter eigener Marke in Umlauf) oder als *Betreiber* (setzt eingekaufte KI ein) auftritt. Diese Unterscheidung ist praktisch relevant, weil sie Umfang und Tiefe von Pflichten, Dokumentation und Verantwortlichkeiten bestimmt. Aus der Einordnung folgen typische Umsetzungsbausteine wie Risikoprüfung und -management, technische Dokumentation und gegebenenfalls Registrierung. Hinzu kommen Transparenzanforderungen, etwa wenn mit KI interagiert wird oder synthetische Inhalte ein Täuschungspotenzial haben, sowie der Aufbau von *AI Literacy*: Mitarbeitende müssen im angemessenen Umfang befähigt werden, KI sachgerecht zu nutzen und Risiken zu erkennen.

Neben der Regulierung bleibt Datenschutz ein Kernbaustein, insbesondere bei automatisierten Entscheidungen mit erheblicher Wirkung. Hier zählen belastbare Rechtsgrundlagen, Datenminimierung, Zweckbindung und Privacy-by-Design zu den wichtigsten Leitplanken. Ergänzend spielen unternehmensrechtliche Anforderungen eine Rolle, etwa der Schutz von Geschäftsgeheimnissen und geistigem Eigentum. Gerade bei externen, generativen KI-Services steigt das Risiko ungewollter Preisgabe sensibler Informationen oder eines Kontrollverlusts über Datenflüsse, Modelle und Updates. Deshalb braucht Responsible AI ein integriertes Governance-Framework, das rechtliche und technische Risikoanalysen, Prüfungen auf Verzerrungen, klare Dokumentations- und Freigabeprozesse, menschliche Aufsicht, Security- und Datenschutzmaßnahmen sowie kontinuierliches Monitoring im Betrieb zusammenführt. KI wird damit nicht „irgendwie“ eingeführt, sondern als steuerbares System mit klaren Zuständigkeiten über den gesamten Lebenszyklus.

Ein weiterer zentraler Hebel ist Erklärbarkeit, weil viele leistungsfähige Modelle von außen schwer nachvollziehbar sind. Erklärbarkeit adressiert das Spannungsfeld zwischen Automatisierung und Verantwortung, indem sie sichtbar macht, wie Ergebnisse zustande kommen und welche Faktoren Entscheidungen beeinflussen. Dabei ist es wichtig, zwischen *Transparenz* (Einblick in Abläufe, Daten und Dokumentation), *Interpretierbarkeit* (von sich aus verständliche Modelle) und *Erklärbarkeit* (Methoden, die Entscheidungen für konkrete Fälle nachvollziehbar aufbereiten) zu unterscheiden. In der Praxis kommen je nach Kontext modellinterne interpretierbare Ansätze, post-hoc Erklärungen für komplexe Modelle sowie datenbezogene Analysen (z. B. Verzerrungen, Ausreißer, Repräsentativität) zum Einsatz. Für Unternehmen ist Explainable AI geschäftsrelevant, weil sie Vertrauen und Akzeptanz erhöht, die Kommunikation zwischen Fachbereichen und Management erleichtert und Haftungs- sowie Qualitätsrisiken senkt. Wirksam wird Erklärbarkeit jedoch erst, wenn Mindeststandards (global, lokal, datenbezogen), saubere Verantwortlichkeiten und eine zielgruppengerechte Aufbereitung zusammenkommen, sodass Erklärungen für Management, Fachbereiche und Prüfinstanzen tatsächlich nutzbar sind.

Mit der breiten Nutzung großer Modelle rückt zudem der Ressourcenverbrauch stärker in den Fokus: Rechenaufwand, Energiebedarf und damit Kosten steigen, insbesondere bei Training und Betrieb. Green AI fordert daher, KI nicht nur nach Leistungskennzahlen, sondern auch nach ökologischen Auswirkungen und Effizienz zu bewerten. Für Unternehmen ist das eine doppelte Chance, weil Effizienzmaßnahmen sowohl Kosten senken als auch Nachhaltigkeitsziele unterstützen. Praktische Hebel sind etwa die Wiederverwendung bestehender Modelle statt Neutraining, ressourcenschonende Adaptionen (Transfer Learning), inkrementelle Updates, Modellkompression für den Betrieb sowie Monitoring und Reporting von Energie- und Ressourceneinsatz. Häufig ist ein pragmatischer Ansatz am wirksamsten: Ein kleineres oder angepasstes Modell kann für den konkreten Prozess ausreichend sein, wenn es robust, wartbar und wirtschaftlich betreibbar ist.

Schließlich sind Architekturentscheidungen ein entscheidender Bestandteil von Responsible AI, insbesondere wenn Datenschutz, Kontrolle, Latenz oder Ausfallsicherheit kritisch sind. Lokale Modelle (On-Premise oder Edge) können Datensouveränität und Resilienz stärken, weil sensible Daten das eigene Netzwerk nicht verlassen müssen und Updates sowie Konfigurationen kontrolliert werden können. Edge-Ansätze verarbeiten Daten direkt am Gerät oder an der Maschine, reduzieren Latenzen, sparen Bandbreite und bleiben auch bei schlechter Verbindung funktionsfähig. Cloud-Lösungen bieten dagegen schnelle Einsatzbereitschaft und Skalierung, verlangen jedoch eine bewusste Steuerung von Datenflüssen und Abhängigkeiten. Für die Domänenanpassung sind insbesondere Fine-Tuning und Retrieval-Augmented Generation (RAG) relevant: Fine-Tuning verankert Unternehmenssprache und Domänenlogik im Modell, während RAG aktuelles, internes Wissen aus Dokumenten und Wissensbasen kontextuell einbindet. In der Praxis ist häufig ein hybrider Ansatz sinnvoll, der Exploration und Skalierungsvorteile der Cloud mit kontrollierten Betriebsformen für sensible, regulierte oder produktionskritische Szenarien kombiniert.

In der Quintessenz entsteht Responsible AI aus der Verknüpfung dieser Dimensionen: Transparenzanforderungen können Erklärbarkeit erzwingen, Nachhaltigkeitsziele beeinflussen Modell- und Betriebsentscheidungen, und Architekturfragen wirken direkt auf Datenschutz, Security und Governance. Für Entscheiderinnen und Entscheider folgt daraus vor allem ein Handlungsprinzip: Nicht einzelne Maßnahmen isoliert optimieren, sondern eine belastbare Bestandsaufnahme, klare Verantwortlichkeiten und einen schrittweisen Umsetzungsplan etablieren, der Rollen, Risiken, Nachweise und Betriebsrealität integriert.

Für Entscheiderinnen und Entscheider liegt der nächste sinnvolle Schritt darin, aus dieser Einordnung eine *Bestandsaufnahme* abzuleiten: Wo wird KI bereits genutzt, welche Risiken und Pflichten entstehen, welche Nachweise fehlen, und welche Zielbilder sind mittelfristig realistisch? Die folgende Liste bündelt dafür konkrete, gut delegierbare Arbeitspakete als **Die nächsten Schritte für Ihr Unternehmen**.

- **KI-Landkarte aufbauen**: Erstellen Sie ein zentrales Register aller KI-Anwendungen (inkl. Schatten-IT): Zweck, Prozessbezug, kritische Entscheidungen, Modell-/Tool-Kategorie, Datenquellen, Schnittstellen, externe Anbieter sowie aktueller Reifegrad (Pilot, Betrieb, geplant).
- **Rolle & Risikoklasse einordnen**: Klären Sie je Use Case, ob Ihr Unternehmen primär als *Anbieter* oder *Betreiber* agiert und wie hoch das Risiko des Einsatzes ist (z. B. hochkritische Entscheidungen vs. unterstützende Automatisierung). Leiten Sie daraus die wichtigsten Nachweis-, Dokumentations- und Kontrollpflichten ab.
- **Governance & Verantwortlichkeiten festlegen**: Definieren Sie klare Zuständigkeiten (z. B. RACI) für KI-Governance, Recht/Compliance, Datenschutz, Informationssicherheit, Fachbereich, Data Science sowie Nachhaltigkeit/ESG. Etablieren Sie Entscheidungspunkte entlang des Lebenszyklus (Beschaffung, Entwicklung, Freigabe, Betrieb, Änderung, Abschaltung).
- **Mindestanforderungen an Transparenz definieren**: Legen Sie fest, *wer* was verstehen muss (Management, Fachbereich, Audit, Betroffene) und welche Erklär- und Nachweisdokumente dafür nötig sind (z. B. Zweckbeschreibung, Datenherkunft, Modellgrenzen, menschliche Aufsicht, Test- & Abnahmeprotokolle).
- **Betrieb absichern**: Planen Sie Monitoring und Incident-Prozesse: Qualität, Drift, Bias-Indikatoren, Sicherheitsvorfälle, Logging, Zugriffskontrolle, Update- & Change-Management sowie Lieferantensteuerung (Verträge, SLAs, Datenflüsse, Exit-Strategien).
- **Nachhaltigkeit & Architektur entscheiden**: Messen bzw. schätzen Sie Ressourcenverbrauch und Kosten wesentlicher KI-Workloads, prüfen Sie Wiederverwendung/Kompression/Fine-Tuning statt Neu-Training und bewerten Sie Cloud vs. On-Prem vs. Edge im Spannungsfeld aus Performance, Datensouveränität, Sicherheit und Energieeffizienz.
- **Roadmap ableiten**: Übersetzen Sie die Befunde in eine priorisierte Roadmap (Quick Wins in 4–8 Wochen, mittelfristige Fähigkeiten in 3–6 Monaten, strategische Zielbilder), inkl. Verantwortlichen, Meilensteinen und klaren Erfolgskriterien.

Literatur

[1] KPMG AG Wirtschaftsprüfungsgesellschaft, *Aus Kür wird Pflicht: 91 Prozent der deutschen Unternehmen sehen KI als geschäftskritisch an und stocken Budgets deutlich auf*, Pressemitteilung, Studie „Generative KI in der deutschen Wirtschaft 2025“, Juni 2025. Adresse: https://kpmg.com/de/de/home/media/press-releases/2025/06/aus-kuer-wird-pflicht-91-prozent-der-deutschen-unternehmen-sehen-ki-als-geschaeftskritisch-an-und-stocken-budgets-deutlich-auf.html (besucht am 05. 09. 2025).

[2] European Commission, *Proposal for a Regulation amending Regulations (EU) 2016/679, (EU) 2018/1724, (EU) 2018/1725, (EU) 2023/2854 and Directives 2002/58/EC, (EU) 2022/2555 and (EU) 2022/2557 as regards the simplification of the digital legislative framework (Digital Omnibus)*, COM(2025) 837 final, 19. Nov. 2025. Adresse: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52025PC0837 (besucht am 10. 01. 2026).

[3] Oberlandesgericht Köln, *Urteil vom 23.05.2025, 15 UKl 2/25 (KI-Training; Art. 6 Abs. 1 lit. f DSGVO; Eilverfahren)*, ECLI:DE:OLGK:2025:0523.15UKL2.25.00, 23. Mai 2025. Adresse: https://nrwe.justiz.nrw.de/olgs/koeln/j2025/15_UKl_2_25_Urteil_20250523.html (besucht am 10. 01. 2026).

[4] OpenAI. „How your data is used to improve model performance.“ (2025), Adresse: https://openai.com/policies/how-your-data-is-used-to-improve-model-performance/ (besucht am 10. 01. 2026).

[5] OpenAI. „Enterprise privacy at OpenAI.“ (2025), Adresse: https://openai.com/enterprise-privacy/ (besucht am 10. 01. 2026).

[6] European Data Protection Supervisor, „Tech Dispatch on Explainable Artificial Intelligence,“ EDPS, Techn. Ber., Nov. 2023, Accessed: 2025-02-14. Adresse: https://www.edps.europa.eu/system/files/2023-11/23-11-16_techdispatch_xai_en.pdf.

[7] European Parliament and the Council of the European Union. „Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 11 (Technical documentation).“ OJ L, 2024/1689, 12.7.2024. See also Annex IV (technical documentation requirements for high-risk AI systems). (13. Juni 2024), Adresse: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng (besucht am 10. 01. 2026).

[8] European Parliament and the Council of the European Union. „Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 13 (Transparency and provision of information to deployers).“ OJ L, 2024/1689, 12.7.2024. Linked transparency obligations for high-risk AI systems. (13. Juni 2024), Adresse: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng (besucht am 10. 01. 2026).

[9] M. Mitchell, S. Wu, A. Zaldivar u. a., „Model Cards for Model Reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*, New York, NY, USA: Association for Computing Machinery, 2019, S. 220–229. Adresse: https://doi.org/10.1145/3287560.3287596 (besucht am 10. 01. 2026).

[10] T. Gebru, J. Morgenstern, B. Vecchione u. a., „Datasheets for Datasets," *Communications of the ACM*, Jg. 64, Nr. 12, S. 86–92, 2021. Adresse: https://doi.org/10.1145/3458723 (besucht am 10. 01. 2026).

[11] C. Molnar, *Interpretable Machine Learning, A Guide for Making Black Box Models Explainable*, 3. Aufl. 2025. Adresse: https://christophm.github.io/interpretable-ml-book (besucht am 10. 01. 2026).

[12] Z. C. Lipton, „The Mythos of Model Interpretability," *Queue*, Jg. 16, Nr. 3, S. 31–57, 2018. Adresse: https://doi.org/10.1145/3236386.3241340 (besucht am 10. 01. 2026).

[13] P. J. Phillips, C. A. Hahn, P. C. Fontana u. a., „Four Principles of Explainable Artificial Intelligence," National Institute of Standards und Technology, NIST Interagency/Internal Report (NISTIR) 8312, 2021. Adresse: https://doi.org/10.6028/NIST.IR.8312 (besucht am 10. 01. 2026).

[14] Y. Rong, T. Leemann, T.-T. Nguyen u. a., „Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Jg. 46, Nr. 4, S. 2104–2122, Apr. 2024. Adresse: http://dx.doi.org/10.1109/TPAMI.2023.3331846.

[15] F. Doshi-Velez und B. Kim, *Towards A Rigorous Science of Interpretable Machine Learning*, 2017. eprint: 1702.08608.

[16] A. Madsen, S. Reddy und A. P. S. Chandar, „Post-hoc Interpretability for Neural NLP: A Survey," *ACM Computing Surveys*, Jg. 55, S. 1–42, 2021.

[17] H. Suresh und J. Guttag, „A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle," in *Equity and Access in Algorithms, Mechanisms, and Optimization*, Ser. EAAMO '21, ACM, Okt. 2021, S. 1–9. Adresse: http://dx.doi.org/10.1145/3465416.3483305.

[18] H. Zhao, H. Chen, F. Yang u. a., *Explainability for Large Language Models: A Survey*, 2023. eprint: 2309.01029.

[19] W. Yang, Y. Wei, H. Wei u. a., „Survey on Explainable AI: From Approaches, Limitations and Applications Aspects," *Human-Centric Intelligent Systems*, Jg. 3, S. 161–188, 2023.

[20] P. Q. Le, M. Nauta, V. B. Nguyen u. a., „Benchmarking eXplainable AI - A Survey on Available Toolkits and Open Challenges," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, E. Elkind, Hrsg., Survey Track, International Joint Conferences on Artificial Intelligence Organization, Aug. 2023, S. 6665–6673. Adresse: https://doi.org/10.24963/ijcai.2023/747.

[21] T. Vermeire, T. Laugel, X. Renard u. a., *How to choose an Explainability Method? Towards a Methodical Implementation of XAI in Practice*, 2021. eprint: 2107.04427.

[22] T. Nguyen, A. Canossa und J. Zhu, *How Human-Centered Explainable AI Interface Are Designed and Evaluated: A Systematic Survey*, 2024. eprint: 2403.14496.

[23] R. Schwartz, J. Dodge, N. A. Smith u. a., „Green AI," *Communications of the ACM*, Jg. 63, Nr. 12, S. 54–63, 2019.

[24] E. Strubell, A. Ganesh und A. McCallum, „Energy and Policy Considerations for Deep Learning in NLP," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, S. 3645–3650.

[25] European Parliament and Council of the European Union, *EU Artificial Intelligence Act*, Final legislative text, 2024.

[26] J. Howard und S. Ruder, „Universal Language Model Fine-tuning for Text Classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, S. 328–339.

[27] M. Tschannen, J. Djolonga, P. K. Rubenstein u. a., *DORA: A Data-Efficient Approach for Fine-Tuning Foundation Models*, 2023. arXiv: 2311.11829.

[28] Green Software Foundation, *Green Software Patterns Catalog*, 2022. Adresse: https://patterns.greensoftware.foundation.

[29] E. J. Hu, Y. Shen, P. Wallis u. a., *LoRA: Low-Rank Adaptation of Large Language Models*, 2021. eprint: 2106.09685.

[30] L. Wang, X. Zhang, H. Su u. a., *A Comprehensive Survey of Continual Learning: Theory, Method and Application*, 2024. arXiv: 2302.00487 [cs.LG]. Adresse: https://arxiv.org/abs/2302.00487.

[31] J. Kirkpatrick, R. Pascanu, N. Rabinowitz u. a., „Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, Jg. 114, Nr. 13, S. 3521–3526, März 2017. Adresse: http://dx.doi.org/10.1073/pnas.1611835114.

[32] S. Grossberg, „How does a brain build a cognitive code?" *Psychological Review*, Jg. 87, Nr. 1, S. 1–51, 1980.

[33] J. O. Neill, *An Overview of Neural Network Compression*, 2020. arXiv: 2006.03669 [cs.LG]. Adresse: https://arxiv.org/abs/2006.03669.

[34] S. Vadera und S. Ameen, *Methods for Pruning Deep Neural Networks*, 2021. arXiv: 2011.00241 [cs.LG]. Adresse: https://arxiv.org/abs/2011.00241.

[35] H. Cheng, M. Zhang und J. Q. Shi, „A Survey on Deep Neural Network Pruning: Taxonomy, Comparison, Analysis, and Recommendations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Jg. 46, Nr. 12, S. 10 558–10 578, 2024.

[36] B. Jacob, S. Kligys, B. Chen u. a., *Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference*, 2017. eprint: 1712.05877.

[37] G. Hinton, O. Vinyals und J. Dean, *Distilling the Knowledge in a Neural Network*, 2015. eprint: 1503.02531.

[38] A. M. Mansourian, R. Ahmadi, M. Ghafouri u. a., *A Comprehensive Survey on Knowledge Distillation*, 2025. arXiv: 2503.12067 [cs.CV]. Adresse: https://arxiv.org/abs/2503.12067.

[39] Google Inc., *Environmental Report 2024*, 2024. Adresse: https://sustainability.google/reports/google-2024-environmental-report/.

[40] Z. Zhang, J. Shi und S. Tang, „Cloud or On-Premise? A Strategic View of Large Language Model Deployment," *SSRN*, 2025.

[41] D. Mart, *On-Premise vs Cloud LLM Hosting — Pros, Cons, and Use Cases*, [Online; accessed 19. Nov. 2025], Nov. 2025. Adresse: https://www.databasemart.com/blog/on-premise-vs-cloud-llm-hosting.

[42] R. Singh und S. S. Gill, „Edge AI: a survey," *Internet of Things and Cyber-Physical Systems*, Jg. 3, S. 71–92, 2023.

[43] S. Trajanoski und A. Karadimce, „Comparative Analysis of Large Language Models: On-Premise Architectures vs. Cloud-Based Deployments," *Preface to Volume 5 Issue 2 of the Journal of University of Information Science and Technology "St. Paul the Apostle"–Ohrid*, Jg. 5, Nr. 2, S. 48, 2025.

[44] Makandra, *Local LLMs in organisations: Using AI securely*, [Online; accessed 19. Nov. 2025], 2025. Adresse: https://makandra.de/en/articles/local-llm-548.

[45] H. Yu, A. Gan, K. Zhang u. a., „Evaluation of retrieval-augmented generation: A survey," in *CCF Conference on Big Data*, Springer, 2024, S. 102–120.

IMPRESSUM

HERAUSGEBER

INHALTLICHE VERANTWORTUNG

Stephan Sandfuchs[1], Diako Farooghi[1], Janis Mohr[2], Sarah Grewe[3], Markus Lemmen[1] und Jörg Frochte[1]
Interdisziplinäres Institut für Angewandte KI und Data Science Ruhr (AKIS)
Bochum University of Applied Sciences

GESCHÄFTSSTELLE TRAIBER.NRW

Bergische Universität Wuppertal
Institute for Technologies and Management for Digital Transformation (TMDT)

Gebäude FZ | Ebene 01 | Raum 19
Lise-Meitner-Str. 27-31, 42119 Wuppertal
Telefon: 0202 439-1164
Email: koordination@traiber.nrw
www.traiber.nrw

Wuppertal, Januar 2026

[1] TrAIBeR.NRW gefördert durch das Bundesministerium für Wirtschaft und Energie unter dem Förderkennzeichen 16TNW0024C
[2] CoFILL gefördert durch das Bundesministerium für Forschung, Technologie und Raumfahrt unter dem Förderkennzeichen 01IS24034B
[3] JetSki gefördert durch das Ministerium für Wirtschaft, Industrie, Klimaschutz und Energie des Landes Nordrhein-Westfalen unter dem Förderkennzeichen EFRE-20800516