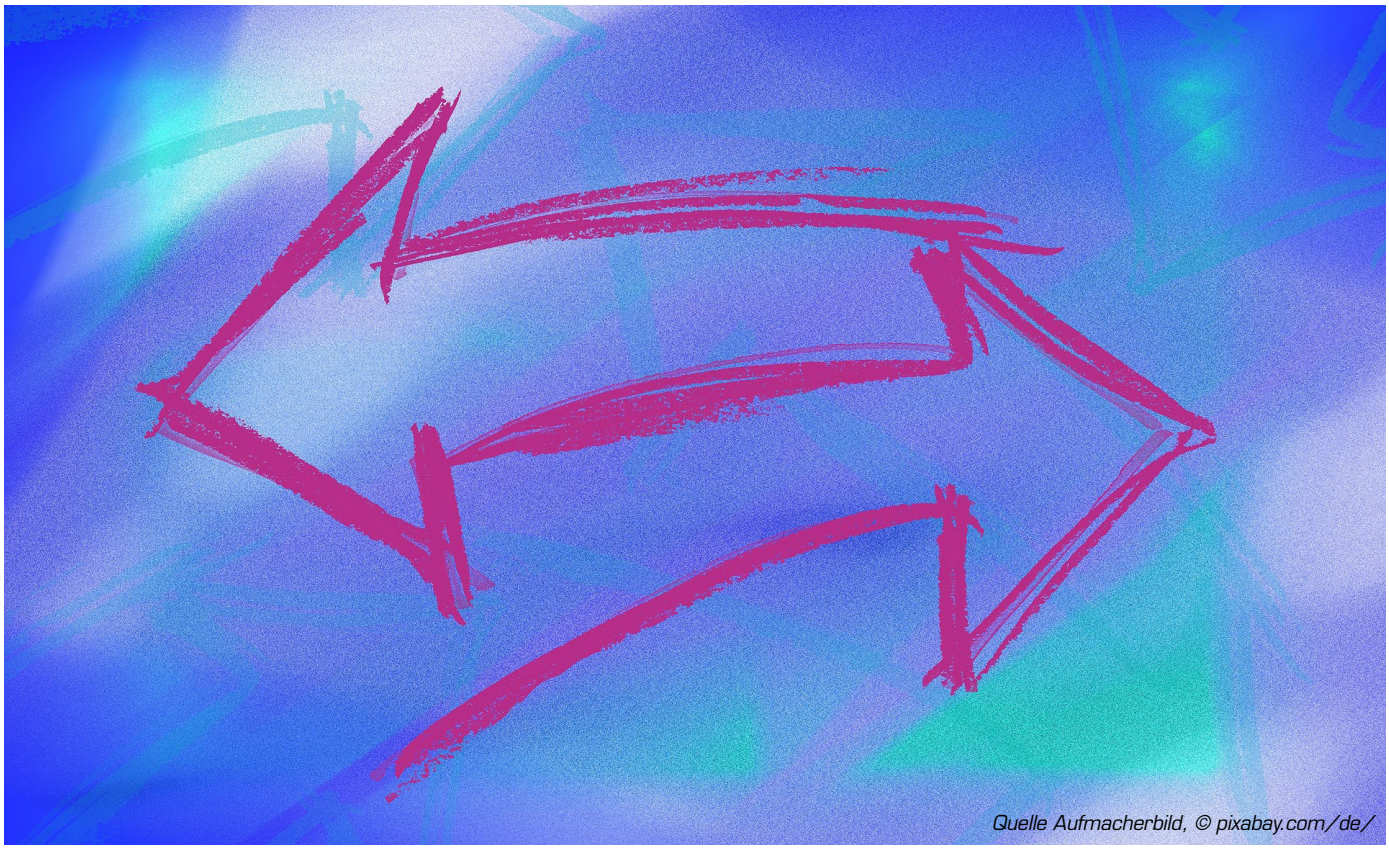
Sonderdruck aus IT Spektrum 03/2026

Ein Erfahrungsbericht über den architektonischen Spagat zwischen generativer KI (RAG) und digitaler Souveränität am Beispiel einer KI-basierten Lösung

Das Schwerpunktthema dieser Ausgabe fragt: „Souverän! Aber wie?“ Für Softwarearchitekten im öffentlichen Sektor ist dies keine philosophische, sondern eine existenzielle Frage. Wir standen vor dem Dilemma, eine hochmoderne KI-Lösung für das komplexe Sozialrecht zu entwickeln, die zwingend auf Hyperscaler-Technologie angewiesen ist, während unsere Kunden – Behörden und Berufsgeheimnisträger – die Nutzung von US-Clouds kategorisch ausschlossen. Dieser Beitrag beschreibt, wie die Architektur der AWS European Sovereign Cloud (ESC) diesen Knoten löste, indem sie operative Autonomie, Metadaten-Residenz und technische Isolation (Nitro System) so kombinierte, dass selbst strengste Compliance-Wächter grünes Licht gaben – ohne dass wir auf externe Hardware-Workarounds zurückgreifen mussten.



Quelle Aufmacherbild, © pixabay.com/de/

Wer heute Softwarearchitekturen für die öffentliche Verwaltung, das Justizwesen oder das Gesundheitswesen in Deutschland entwirft, bewegt sich in einem Spannungsfeld der Reaktion auf das „wenig vorhersagbare Verhalten“ globaler Akteure.

Wenn Innovation auf Realpolitik trifft

Auf der einen Seite steht der immense *Druck der Fachbereiche*. Der Fachkräftemangel in den Amtsstuben ist real, die Aktenberge wachsen, und der Ruf nach Entlastung durch Automatisierung und Künstliche Intelligenz (KI) wird lauter. Die Erwartungshaltung ist klar: Die IT soll Lösungen liefern, die so intuitiv und mächtig sind wie ChatGPT, aber so sicher wie ein Tresor im Keller des Bundeskanzleramts.

Auf der anderen Seite steht die *Compliance-Mauer*. Datenschutzgrundverordnung (DSGVO), Sozialdatenschutz (§ 67 SGB X) und das strafbewehrte Berufsgeheimnis (§ 203 StGB) bilden einen festen Rahmen. Doch das eigentliche Hindernis ist oft nicht der Gesetzestext selbst, sondern die geopolitische Risikobewertung. Die „Digitale Souveränität“ ist zur harten Währung geworden. Sie definiert sich durch die Fähigkeit, selbstbestimmt über digitale Infrastrukturen zu verfügen, ohne einseitige Abhängigkeiten oder die Gefahr der Einflussnahme durch Drittstaaten – Stichwort: US CLOUD Act.

In diesem Szenario entwickelten wir einen KI-Assistenten (msg.CodIQ | msg) für das deutsche Sozialgesetzbuch (SGB). Unsere

Reise von der technischen Machbarkeit hin zur akzeptierten Lösung ist ein Lehrstück darüber, wie sich das scheinbare Paradoxon aus „Amerikanischer High-Tech“ und „Europäischer Souveränität“ auflösen lässt.

Der Anwendungsfall: Grenzen von On-premises und warum wir die Cloud brauchten

Um die architektonische Brisanz zu verstehen, muss man zunächst das fachliche Problem betrachten. Das deutsche Sozialrecht (SGB I bis XIV) ist eines der komplexesten Regelwerke weltweit. Es ist ein lebender Organismus, der sich durch Gesetzesnovellen, Länderverordnungen und täglich neue Urteile der Sozialgerichte ständig verändert. Für Sachbearbeitende in Jobcentern oder Fachanwälte ist es fast unmöglich, ohne intelligente Hilfsmittel den Überblick zu behalten.
Unser Ansatz war die Entwicklung einer RAG-Anwendung (Retrieval-Augmented Generation). Im Gegensatz zu simplen Chatbots, die nur auf trainiertem Wissen basieren und zu Halluzinationen neigen, sucht unser System zuerst faktenbasierte Quellen (Gesetze, Urteile), um diese dann einem Large Language Model (LLM) zur Verarbeitung vorzulegen.
Technisch erfordert dies eine Kette von Hochleistungsservices:

1. *Semantische Vektorisierung:* Zerlegung von Tausenden Gesetzesseiten in semantische Vektoren (Embeddings). Hierbei müssen Modelle verwendet werden, die das komplexe „Juristendeutsch“ nicht nur übersetzen, sondern in seinen Nuancen verstehen.
2. *Vektorsuche:* Eine Datenbank, die Millisekunden-schnelle Ähnlichkeitssuche im hochdimensionalen Raum ermöglicht.
3. *Inferenz:* Zugriff auf leistungsstarke Foundation Models, die riesige Kontextfenster verarbeiten können, um ganze Fallakten zu analysieren.

Schnell wurde klar: In einem klassischen On-premises-Rechenzentrum einer Behörde ist dies kaum wirtschaftlich abbildbar. Die Beschaffung von GPU-Clustern für die Inferenz, der Betrieb von hochverfügbaren Vektordatenbanken (wie OpenSearch) und das ständige Aktualisieren der LLMs erfordern eine Agilität, die nur Cloud-Plattformen bieten. Wir brauchten die Skalierbarkeit und den Managed-Service-Ansatz eines Hyperscalers, um die Anwendung kosteneffizient und performant anbieten zu können.

Die Blockade: Das „Nein“ der Kunden verstehen

Als wir mit dem Prototypen unserer Lösung auf den Markt gingen, erlebten wir eine Dichotomie. Fachlich rannten wir offene Türen ein. Doch sobald das Wort „Public Cloud“ oder gar „AWS“ fiel, schlossen sich diese Türen wieder.
Die Ablehnung war nicht emotional, sondern rational begründet. Wir führten intensive Gespräche mit IT-Sicherheitsbeauftragten der Länder und Kanzleien. Deren Bedenken ließen sich auf drei Kernpunkte kondensieren, die jede Architektur adressieren muss, die im Jahr 2026 „souverän“ sein will:

- *Die extraterritoriale Zugriffsmöglichkeit:* Selbst wenn Daten verschlüsselt in Frankfurt liegen: Solange der Betreiber ein US-Unternehmen ist, unterliegt er dem US CLOUD Act oder FISA (Foreign Intelligence Surveillance Act). Die Sorge war, dass ein US-Gericht den Anbieter zwingen könnte, Daten herauszugeben – vorbei an europäischen Rechtshilfeverfahren. Für Berufsgeheimnisträger ist dieses theoretische Risiko oft ein Show-Stopper.
- *Das Metadaten-Problem:* In der juristischen Arbeit sind Metadaten oft so brisant wie die Inhaltsdaten. Zu wissen, dass eine bestimmte Kanzlei um 3 Uhr nachts massiv Datenbankabfragen zum Thema „Insolvenzverschleppung“ stellt, ist ein verräterisches Datum. In klassischen Cloud-Architekturen ist die „Control Plane“ (Identitätsmanagement, Billing, Logging) oft global vernetzt. Metadaten fließen zur Abrechnung in die USA ab. Für unsere Kunden war das ein „No-Go“.
- *Operationale Abhängigkeit:* Souveränität bedeutet auch, handlungsfähig zu bleiben, wenn das transatlantische Kabel gekappt wird – sei es physisch oder politisch. Ein System, das für Wartungsarbeiten auf Administratoren in Seattle oder Support-Teams in Indien angewiesen ist (Follow-the-Sun-Modell), gilt als nicht souverän.

Wir standen vor einer klassischen „Build vs. Buy“-Sackgasse. Bauen wir alles selbst in einem deutschen Rechenzentrum und verlieren den technologischen Anschluss? Oder nutzen wir den Hyperscaler und verlieren die Kunden?

Der architektonische Ausweg: AWS European Sovereign Cloud

In dieser Phase evaluierten wir die neu angekündigte *AWS European Sovereign Cloud (ESC)*. Für uns als Architekten war entscheidend, hinter die Marketing-Fassade zu blicken. Handelt es sich nur um „Region Frankfurt mit neuem Sticker“ oder um eine fundamentale Änderung der Topologie?
Unsere Analyse zeigte, dass die ESC genau jene Architekturmuster implementiert, die zur Auflösung unseres Dilemmas notwendig waren. Wir haben unsere Lösung darauf migriert, weil die ESC drei spezifische „Trust Boundaries“ (Vertrauensgrenzen) verschiebt: die menschliche Firewall, die Daten-Firewall und die technische Firewall.

Die menschliche Firewall: operative Autonomie

Das stärkste Argument der Cloud-Skeptiker war immer der „Rogue Admin“ oder der erzwungene Zugriff durch Support-Personal. Die ESC implementiert hier eine harte Trennung. Sie wird nicht von Amazon Web Services, Inc. (USA) betrieben, sondern von einer eigenständigen juristischen Person in der EU:

- *Residency-Zwang:* Administratoren, die physischen oder logischen Zugang zur Infrastruktur haben, müssen in der EU ansässig sein und bei der EU-Gesellschaft angestellt sein. Sie unterliegen EU-Recht.
- *Kein Remote-Zugriff:* Es gibt technisch keine VPN-Tunnel oder Wartungsschnittstellen, die es einem Engineer aus den USA erlauben würden, auf die Systeme der ESC zuzugreifen.

Für unsere Risikoanalyse für unsere Lösung bedeutete das: Der Vektor „Zwang durch US-Behörden“ ist deutlich erschwert. Selbst wenn die US-Muttergesellschaft unter Druck gesetzt würde, Daten herauszugeben, ist ihr der technische Zugriffspfad (Access Control im Messaging and Positioning Framework/MPF) stark eingeschränkt. Dies ist ein belastbares Argument gegenüber unseren Kunden.

Die Daten-Firewall: Metadaten-Residenz und Autarkie

Als Architekten mussten wir uns intensiv mit dem Unterschied zwischen *Data Plane* und *Control Plane* auseinandersetzen. In der Standard-Region eu-central-1 bleibt zwar mein S3-Bucket in Frankfurt (Data Plane), aber mein IAM-User oder meine CloudTrail-Logs (Control Plane) werden unter Umständen global repliziert.
Die ESC ist eine *isolierte Cloud*. Sie teilt sich keine Control Plane mit den globalen Regionen:

- Sie hat ein eigenes Identity and Access Management (IAM).
- Sie hat ein eigenes Billing-System.
- Sie hat eigene Metering-Logs.

Das bedeutet: Wenn ein Anwalt unsere Lösung nutzt, verlassen weder die Fallakten (Content) noch die Informationen über seine Nutzung (Metadata) den europäischen Rechtsraum. Dies war der Schlüssel, um die strengen Anforderungen des Datenschutzes bei Sozialdaten zu erfüllen.
Mehr noch: Die Cloud ist „disconnected capable“. Sie ist so gebaut, dass sie auch ohne Verbindung zum globalen Internet oder dem AWS-Backbone voll funktionsfähig bleibt. Dies erhöht die *Resilienz* kritischer staatlicher Anwendungen massiv.

Die technische Firewall: das Nitro System

Verträge sind gut, technische Garantien sind besser. Ein Highlight unserer Argumentation gegenüber technikaffinen Kunden (z. B. den IT-Dienstleistern der Justiz) ist das *AWS Nitro System*. Bei einer RAG-Anwendung müssen Daten (z. B. die Fallakte) kurzzeitig unverschlüsselt im Arbeitsspeicher (RAM) der EC2-Instanz liegen, damit das LLM sie verarbeiten kann. Die klassische Angst: „Der Cloud-Provider kann den RAM auslesen.“
Das Nitro System entkräftet dies durch Hardware-Design (Sovereign by Design). Die Virtualisierung wurde auf dedizierte Chips ausgelagert. Der Hypervisor ist so minimalistisch, dass er keine administrative Konsole (wie SSH) bietet. Es gibt für einen AWS-Mitarbeiter keinen Befehl, um den Speicher einer Kundeninstanz zu dumpen. Dies ist ein architektureller Schutz, der unabhängig von menschlichem Handeln greift. Er wurde von unabhängigen Stellen (wie der NCC Group) auditiert und bestätigt.

Umsetzung in der Praxis: der „Sovereign-native“-Ansatz

Die Entscheidung für die ESC war strategisch richtig, brachte aber auch neue Herausforderungen in der Implementierung mit sich. Wer eine souveräne Architektur baut, muss einige Bequemlichkeiten der globalen Cloud überdenken.

Vertrauen in die Plattform statt externer Workarounds

Oft wird in Souveränitätsdiskussionen gefordert, Schlüsselmaterial zwingend außerhalb der Cloud zu halten (External Key Stores/HYOK). Wir haben für unsere Lösung nach eingehender Analyse bewusst entschieden, primär auf die *native Verschlüsselung der ESC (AWS KMS)* zu setzen.
Warum? Weil die ESC selbst die Vertrauensgarantie liefert. Da der Key Management Service (KMS) in der ESC vollständig isoliert ist und von EU-Personal betrieben wird, verlassen die Schlüssel niemals den europäischen Rechtsraum und sind dem Zugriff der US-Muttergesellschaft entzogen. Dieser „Managed Sovereignty“-Ansatz hat entscheidende Vorteile gegenüber externen Key Stores:

- *Performance:* Wir vermeiden die Latenzzeiten, die entstehen, wenn für jede Entschlüsselung ein externes HSM angefragt werden muss. Für eine Echtzeit-KI-Anwendung ist dies kritisch.
- *Verfügbarkeit:* Wir müssen uns nicht um die Hochverfügbarkeit eigener HSM-Cluster kümmern. Die Resilienz der Cloud-Plattform greift auch für das Schlüsselmanagement.

Wir argumentieren gegenüber unseren Kunden so: Da die ESC bereits operativ und technisch (durch Nitro) isoliert ist, ist das Vertrauen in den Cloud-internen KMS-Service gerechtfertigt. Es ist eine Abwägung zwischen extremer Paranoia und operativer Exzellenz – und die ESC erlaubt es uns, diese Balance zugunsten der Exzellenz zu verschieben, ohne die Souveränität zu opfern.

Herausforderung: Service Parity

Ein Lernprozess war der Umgang mit der Verfügbarkeit von Services. Eine isolierte Cloud bedeutet, dass neue Features nicht zeitgleich mit der Region us-east-1 (Nord-Virginia) verfügbar sind. Jeder Service (z. B. neue Bedrock-Modelle) muss erst für die ESC portiert und auditiert werden.
Wir mussten unsere Architektur modular aufbauen. Für unsere Lösung prüfen wir, welche KI-Modelle in der ESC verfügbar sind. Wir können nicht blindlings die neuesten Beta-Features aus den USA einbauen, sondern müssen die Roadmap der ESC genau verfolgen. Das erfordert ein disziplinierteres Release-Management als in der „Wild West“-Welt der globalen Public Cloud.

Kosten und Transparenz

Souveränität gibt es nicht zum Nulltarif. Der Betrieb einer dedizierten Infrastruktur mit lokalem Personal ist kostenintensiver als die Nutzung globaler Ressourcenpools. Wir haben gelernt, dass Transparenz hier entscheidend ist. Kunden im öffentlichen Sektor sind bereit, diesen Aufpreis zu zahlen, wenn sie verstehen, dass er der Preis für ihre Unabhängigkeit und Rechtssicherheit ist. Wir verkaufen nicht mehr nur „Software“, sondern „konforme Softwarearchitektur“.

Vergleich: Warum nicht Open Source oder EU-Cloud-Anbieter?

Warum haben wir unsere Lösung nicht auf einer reinen EU-Cloud (z. B. auf Basis von OpenStack) oder bei einem deutschen Hoster gebaut?
Die Antwort liegt in der *Innovationsgeschwindigkeit*. Wir konkurrieren mit Lösungen, die global entwickelt werden. Ein rein deutscher Anbieter kann unmöglich das gleiche Investitionsvolumen in die Entwicklung von Vektordatenbanken, Serverless-Infrastrukturen und KI-Beschleunigern stecken wie ein globaler Hyperscaler. AWS investiert Milliarden in die ESC – Summen, die lokale Anbieter kaum aufbringen können.
Hätten wir auf einen lokalen Nischenanbieter gesetzt, hätten wir technologische Kompromisse eingehen müssen:

- *Langsamere Inferenzzeiten:* Fehlende optimierte GPU-Instanzen.
- *Schlechtere Skalierbarkeit:* Probleme bei Lastspitzen (z. B. nach einer Gesetzesreform beim Bürgergeld, wenn Tausende Sachbearbeiter gleichzeitig zugreifen).
- *Höherer Wartungsaufwand:* Wir hätten Datenbank-Cluster selbst betreiben müssen, statt Managed Services zu nutzen.

Die AWS ESC ist für uns der „Sweet Spot“. Sie bietet die APIs und die Power des Weltmarktführers (Full Power of Cloud), aber unter den *rechtlichen und operativen Rahmenbedingungen* eines europäischen Anbieters. Es ist der Versuch, das Beste aus beiden Welten zu vereinen – und aus unserer Sicht der derzeit einzige Weg, um anspruchsvolle KI-Workloads im regulierten Umfeld produktiv zu setzen.

Fazit: Souveränität als Wettbewerbsvorteil

Die digitale Souveränität wird oft als Bremsklotz der Digitalisierung wahrgenommen. Unsere Erfahrung bei der Entwicklung unserer KI-basierten Lösung zeigt das Gegenteil: Sie ist ein Enabler. Ohne die harte architektonische Zusicherung der Souveränität hätten wir im deutschen

Referenzen und weiterführende Literatur

AWS Digital Sovereignty Messaging and Positioning Framework (MPF): Grundlegendes Dokument zum Verständnis der vier Säulen (Data Residency, Granular Access, Encryption, Resilience)

Strafgesetzbuch (StGB) § 203: Verletzung von Privatgeheimnissen – Die rechtliche Hürde für Berufsgeheimnisträger in der Cloud, siehe: **https://www.gesetze-im-internet.de/stgb/__203.html**

AWS Nitro System Security Design: Whitepaper zur Hardware-basierten Isolation und dem Fehlen von administrativen Zugriffen (Operator Access), siehe: **https://docs-aws.amazon.com.rproxy.goskope.com/whitepapers/latest/security-design-of-aws-nitro-system/security-design-of-aws-nitro-system.html**

Schrems II (EuGH C-311/18): Das Urteil, das die Anforderungen an den Datentransfer in Drittländer neu definierte und den Bedarf an souveränen Clouds zementierte, siehe: **https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:62018CJ0311**

msg.CodIQ | msg: Detaillierte Informationen zur RAG-Architektur im Sozialrecht und den spezifischen Herausforderungen der Halluzinationsvermeidung, siehe: **https://www.msg.group/de/branchen/public-sector/loesungen-fuer-die-digitale-verwaltung/msg-codiq**

öffentlichen Sektor keine Chance gehabt, eine KI-Lösung einzuführen, die sensible Bürgerdaten verarbeitet. Die Diskussionen wären im Keim erstickt.
Durch die Nutzung der AWS European Sovereign Cloud konnten wir die Diskussion von einer emotionalen („Wir vertrauen den Amerikanern nicht“) auf eine faktische Ebene („Hier ist der technische Beweis, dass kein Zugriff möglich ist“) heben.
Wir haben gelernt:

- *Souveränität ist mehrschichtig:* Es geht nicht nur um den Standort der Daten, sondern um Metadaten, operative Kontrolle und Software-Supply-Chain.
- *Technologie schlägt Vertrag:* Das Nitro System ist ein stärkeres Argument als jede AGB-Klausel.
- *Kein Kompromiss bei der Leistung:* Öffentliche Verwaltung darf nicht bedeuten, dass man mit der Technik von gestern arbeitet.

Die Botschaft ist klar: Wir müssen uns nicht mehr zwischen Innovation und Datenschutz entscheiden. Die neuen Cloud-Betriebsmodelle erlauben es uns, souverän zu bleiben, ohne uns zu isolieren.

Die Autoren

Richard Pielczyk
richard.pielczyk@msg.group
www.linkedin.com/in/richard-pielczyk-65094770

ist seit März 2016 bei msg Public Sector im Bereich Arbeit und Soziales tätig. Seine Haupttätigkeitsfelder liegen in der Entwicklung und Implementierung von praxisnahen Methoden zur Gestaltung von Projekten und Produkten, mit dem Schwerpunkt Sozialwirtschaft. Er arbeitet seit 38 Jahren im Management von IT-Projekt- und Linienorganisationen unterschiedlicher Größenordnung.

Dr.-Ing. Pascal Hinrichs
pascal.hinrichs@msg.group
www.linkedin.com/in/dr-ing-pascal-hinrichs-55758529b
https://orcid.org/0000-0003-3244-715X

ist seit April 2024 bei msg Public Sector im Bereich Arbeit und Soziales tätig. Dort gestaltet er innovative Lösungen für die Sozialwirtschaft. Zuvor promovierte er im Bereich Robotik und KI mit Fokus auf die ambulante Pflege. Sein Schwerpunkt liegt auf der Entwicklung von Systemen, die Fachkräfte vor Ort bei ihrer täglichen Arbeit unterstützen und spürbar entlasten.