Synapse Delić

ENTERPRISE SOVEREIGN AI · PRAXISLEITFADEN 2026

Voice AI & Edge-Inferenz im Mittelstand

DSGVO, EU AI Act und praxisnahe Zero-UI-Architekturen für den operativen Kern

Emil Delić
Quality, Security & Safety Management · Gründer Synapse Delić

Synapse Delić UG (haftungsbeschränkt) · Gifhorn
Ausgabe September 2026 · mail@synapse-delic.de · synapse-delic.de

Inhaltsverzeichnis

Management Summary & Vorwort 3
Über den Autor: Quality, Security & Safety Management 4
Kapitel 1: DSGVO und EU AI Act im Praxiseinsatz 5
1.1 Das rechtliche Fundament: Warum der EU AI Act die DSGVO nicht heilt 5
1.2 Die drei tödlichen Haftungsfallen für mittelständische Betriebe 6
1.3 Der Betriebsrat als Gatekeeper (§ 87 Abs. 1 Nr. 6 BetrVG) 7
1.4 Risikoklassifizierung nach dem EU AI Act im Betrieb 8
1.5 Der architektonische Befreiungsschlag: 100 % Lokale Edge-Inferenz 8
1.6 Praxis-Checkliste vor dem KI-Rollout 8
1.7 Quellen & Rechtsprechungs-Referenzen 9
Kapitel 2: Vom Rechenzentrum ins Funkloch: Das technische How-To 10
2.1 Die 4-Stufen-Schrumpf-Pipeline: Von 140 GB auf 500 MB 10
2.2 Die Zero-UI-Systemarchitektur: Voice-to-JSON Pipeline 10
2.3 Hardware-Dimensionierung für den Mittelstand 11
Kapitel 3: Das Synapse-Delić-Vorgehensmodell 12
3.1 Der 14-Tage High-Speed Zero-UI MVP 12
3.2 Software-Ownership statt SaaS-Abofalle 12
Kapitel 4: Abschluss & Ausblick: Der Weg zur betrieblichen KI-Souveränität 13
4.1 Ausblick 2026/2027: Die Evolution der Edge-Intelligenz 13

SYNAPSE DELIĆ · PRAXISBUCH

Management Summary & Vorwort

Künstliche Intelligenz hat das Potenzial, den deutschen Mittelstand von seiner größten operativen Last zu befreien: Manueller Zettelwirtschaft, quälend langsamen ERP-Eingabemasken und dem massiven Informationsverlust zwischen Baustelle, Werkshalle oder LKW und der kaufmännischen Verwaltung.

Dennoch scheitern im Jahr 2026 über 80 Prozent aller betrieblichen KI-Initiativen. Der Grund dafür liegt selten in der Technologie selbst, sondern an zwei fundamentalen Fehlannahmen:

1. **Die rechtliche Naivität:** Unternehmen verlassen sich gutgläubig auf US-amerikanische Cloud-APIs (wie OpenAI oder Microsoft Copilot) und Standard-Auftragsverarbeitungsverträge. Dabei übersehen sie die juristische Realität: Den BGH-Schadensersatz bei bloßem Datenkontrollverlust, die zwingende Mitbestimmung des Betriebsrats bei Stimmanalysen und die Gefahr, nach dem EU AI Act unbemerkt zum voll haftenden System-Hersteller („frisiertes Mofa“) zu werden.
2. **Die physikalische Ignoranz:** Rechenzentrums-Modelle mit 70 Milliarden Parametern verlangen permanente High-Speed-Netze. In der realen Wirtschaft – im Stahlbetonkeller, im Tunnel oder auf Landstraßen – herrscht jedoch das Funkloch.

Dieses Whitepaper beschreibt den Gegenentwurf: **Sovereign Edge AI**.

Wir zeigen detailliert auf, wie moderne Sprachmodelle durch *Knowledge Distillation* und *INT4-Quantisierung* so komprimiert werden, dass sie in unter 300 Millisekunden zu 100 Prozent autark auf mobilen Endgeräten rechnen – ohne dass ein einziges Byte das Gerät verlässt. Damit wird Datenschutz vom Innovationshemmnis zum schärfsten operativen Wettbewerbsvorteil.

Emil Delić
Gifhorn, im September 2026

SYNAPSE DELIĆ · PRAXISBUCH

Über den Autor: Quality, Security & Safety Management

Emil Delić
Gründer Synapse Delić

Wer heute über Künstliche Intelligenz im Mittelstand spricht, kommt fast immer aus einer von zwei getrennten Welten: Entweder aus der theoretischen Welt der Beratungshäuser und Rechenzentren – oder aus dem harten operativen Tagesgeschäft auf der Baustelle, in der Werkshalle oder im LKW.

Emil Delić verbindet ingenieurwissenschaftliche Präzision mit den realen Anforderungen der betrieblichen Praxis:

Nach seinem Master-Abschluss in Elektronik und Automatisierung spezialisierte er sich konsequent auf Quality-, Security- und Safety-Management in sicherheitskritischen Umgebungen. Als zertifizierter Functional Safety Manager (ISO 26262 / IEC 61508) begleitete und entwickelte er Embedded-Systeme in der Hochtechnologie – dort, wo Softwareausfälle keine Option sind und Systeme auch im totalen Netzausfall fehlerfrei, deterministisch und in Echtzeit reagieren müssen.

Aus der engen Begleitung operativer Betriebe im Handwerk, in der Bauwirtschaft und im technischen Service kennt er gleichzeitig den allabendlichen Frust über verlorene Notizzettel, unleserliche Regieberichte und quälend langsame Softwaremasken aus nächster Nähe.

Aus dieser Synthese entstand **Synapse Delić**: Ein Technologieunternehmen mit Sitz in Gifhorn, das keine bunten Folien malt, sondern praxiserprobte Zero-UI-Sprach-KI baut. Kompakte Modelle, die das gesprochene Wort von Mitarbeitern vor Ort direkt auf dem Smartphone im flüchtigen Speicher in VOB-konforme, prüffähige Daten verwandeln – zu 100 Prozent offline, ohne Datenabfluss und ohne SaaS-Abofalle.

„Ich habe dieses Whitepaper geschrieben, um Geschäftsführern und IT-Leitern einen ehrlichen, ungeschminkten Leitfaden an die Hand zu geben: Wie Sie Ihre Prozesse mit Sprach-KI radikal beschleunigen, ohne in die tödlichen Haftungsfallen der US-Clouds oder die Blockade des Betriebsrats zu laufen.“

SYNAPSE DELIĆ · PRAXISBUCH

Kapitel 1: DSGVO und EU AI Act im Praxiseinsatz

> **Praxisfall: Die 100.000-€-Haftungsfalle der „Bau & Montage Nord GmbH“**
>
> *Geschäftsführer Thomas wollte die Zettelwirtschaft beenden und führte für seine 35 Bauleiter gutgläubig einen modernen Cloud-Sprachassistenten eines US-Anbieters ein. Vier Monate später eskalierte die Situation: Der Betriebsrat schaltete einen Fachanwalt ein, weil die Sprachaufnahmen im US-Rechenzentrum gespeichert wurden und Leistungsanalysen ermöglichten (§ 87 Abs. 1 Nr. 6 BetrVG). Gleichzeitig forderte ein gekündigter Mitarbeiter gestützt auf das BGH-Urteil zum Kontrollverlust 2.500 Euro Schadensersatz nach Art. 82 DSGVO. Und weil die IT das System per Prompting zur automatisierten Mängelbewertung erweitert hatte, drohte die Aufsichtsbehörde mit Einstufung als Hochrisiko-System nach Art. 25 AI Act („frisiertes Mofa“). Der Rollout wurde mit sofortiger Wirkung gestoppt – Gesamtschaden: über 85.000 Euro und monatelanger Stillstand.*

In der betrieblichen Praxis herrscht ein weit verbreiteter und hochgradig haftungsträchtiger Irrglaube: Viele Unternehmen gehen davon aus, dass ein KI-System, welches die Vorgaben der europäischen KI-Verordnung (EU AI Act) erfüllt oder mit einer CE-Kennzeichnung versehen ist, automatisch auch datenschutzrechtlich zulässig sei.
Wie führende IT- und Datenschutzjuristen – unter anderem in der fundierten Analyse des renommierten c’t-Datenschutz-Podcasts *„Auslegungssache“* (Folge 132: *„KI-Verordnung und Datenschutz – ein schwieriges Verhältnis“* mit Prof. Dr. Rolf Schwartmann und Fachanwalt Joerg Heidrich) – herausarbeiten, bewegen sich Unternehmen beim Einsatz von Künstlicher Intelligenz in einem strikten dualen Rechtsregime:

1. **Der EU AI Act ist Produkt- und Sicherheitsrecht:** Er reguliert das Inverkehrbringen, die Inbetriebnahme und die Verwendung von KI-Systemen als Produkte auf dem EU-Binnenmarkt. Sein Ziel ist die Gewährleistung von Produktsicherheit, Nichtdiskriminierung und der Schutz von Grundrechten vor Fehlfunktionen der Technologie. In Deutschland wird die Marktüberwachung hierfür maßgeblich bei der Bundesnetzagentur (BNetzA) angesiedelt.
2. **Die DSGVO ist Verarbeitungs- und Abwehrrecht:** Sie schützt das Grundrecht auf informationelle Selbstbestimmung natürlicher Personen bei der Verarbeitung personenbezogener Daten. Die Aufsicht obliegt den unabhängigen Datenschutzaufsichtsbehörden der Länder und des Bundes (Datenschutzkonferenz, DSK).

1.1 Das rechtliche Fundament: Warum der EU AI Act die DSGVO nicht heilt

Die Unberührt-Klausel (Art. 2 Abs. 7 Satz 2 AI Act)

Der Gesetzgeber hat das Verhältnis beider Gesetze in Art. 2 Abs. 7 Satz 2 AI Act unmissverständlich klargestellt: Das Datenschutzrecht der Europäischen Union bleibt durch die KI-Verordnung unberührt. Der AI Act ist ausdrücklich **keine** *Lex specialis*, die datenschutzrechtliche Vorgaben verdrängt oder suspendiert.
Das bedeutet für jedes Unternehmen:

- Ein KI-System kann technisch zu 100 Prozent konform mit dem AI Act sein und dennoch wegen eines Verstoßes gegen Art. 6 Abs. 1 DSGVO (fehlende Rechtsgrundlage) oder Art. 44 ff. DSGVO (unzulässiger Drittstaatentransfer) von der Datenschutzaufsicht stillgelegt und mit existenzbedrohenden Bußgeldern belegt werden.
- Der AI Act enthält **keine eigene Rechtsgrundlage** für die Verarbeitung personenbezogener Daten im laufenden Betrieb.

Die einzigen zwei Brücken zwischen AI Act und DSGVO

Nur an zwei eng umgrenzten Stellen berühren sich beide Gesetzeswerke direkt:

- **Art. 10 Abs. 5 AI Act (Bias-Korrektur):** Erlaubt Anbietern von Hochrisiko-Systemen ausnahmsweise die Verarbeitung besonderer Kategorien personenbezogener Daten (Art. 9 Abs. 1 DSGVO – z. B. ethnische Herkunft, Gesundheitsdaten), sofern dies zwingend erforderlich ist, um Verzerrungen (Bias) im System zu erkennen und zu korrigieren.
- **Art. 59 Abs. 1 AI Act (Reallabore):** Erlaubt die Weiterverarbeitung von rechtmäßig erhobenen Daten für andere Zwecke innerhalb behördlich überwachter KI-Reallabore (Sandboxes) unter strengen Auflagen gemäß Art. 6 Abs. 4 DSGVO.

Außerhalb dieser beiden eng definierten Spezialfälle muss jede KI-gestützte Verarbeitung personenbezogener Daten die volle Härte der DSGVO eigenständig bestehen.

1.2 Die drei tödlichen Haftungsfallen für mittelständische Betriebe

Unternehmen, die gängige Cloud-KI-Modelle oder Allzweck-Sprachmodelle (General Purpose AI / GPAI) in ihre Betriebsabläufe integrieren, laufen blind in drei gravierende Haftungsfallen:

Falle A: Das „frisierte Mofa" (Die fatale Zweckänderung nach Art. 25 AI Act)

Die meisten mittelständischen Unternehmen betrachten sich beim Einsatz fremder KI-Modelle (wie OpenAI GPT-4, Microsoft Azure AI oder Google Cloud Vertex) als reine **Betreiber (Deployer)**. Für Betreiber gelten im Wesentlichen Verwendungs- und Aufsichtspflichten.

Die juristische Gefahr der Zweckänderung wird von Medienrechtler Prof. Dr. Rolf Schwartmann (GDD-Vorstandsvorsitzender) treffend mit dem Bild des **„frisierten Mofas"** veranschaulicht: Wer ein harmloses Serien-Mofa technisch so modifiziert, dass es 80 km/h fährt, verliert den Schutz der ursprünglichen Betriebserlaubnis und haftet wie der Hersteller eines Motorrads – inklusive aller Führerschein-, Zulassungs- und Sicherheitsauflagen.

Genau das passiert im KI-Recht: Wird ein Allzweck-Sprachmodell (GPAI) über eigene System-Prompts, Programmierschnittstellen (APIs) oder Fine-Tuning in einen Bereich überführt, der unter Anhang III des AI Acts als **Hochrisiko** eingestuft ist (beispielsweise automatisierte Vorauswahl von Stellenbewerbern, Bewertung von Mitarbeiterleistungen, Bonitätsprüfungen oder Steuerung kritischer Infrastruktur), greift Art. 25 AI Act:

- Wer den Verwendungszweck eines KI-Systems wesentlich verändert, verliert seinen Status als bloßer Betreiber und **wird kraft Gesetzes zum Anbieter (Provider / Hersteller)** des Systems.
- **Die Rechtsfolge:** Das mittelständische Unternehmen haftet plötzlich vollumfänglich wie der ursprüngliche Entwickler. Es muss ein vollständiges Konformitätsbewertungsverfahren durchlaufen, ein permanentes Risikomanagementsystem nach Art. 9 AI Act etablieren, technische Dokumentationen vorhalten, das CE-Zeichen anbringen und die Post-Market-Überwachung organisieren.
- Für ein mittelständisches Unternehmen ist dieser administrative und finanzielle Aufwand nicht leistbar. Die zivil- und haftungsrechtliche Verantwortung bei Fehlfunktionen liegt zu 100 Prozent bei der Geschäftsführung.

Falle B: Die BGH-Kontrollverlust-Rechtsprechung (Art. 82 DSGVO)

Lange Zeit galt im deutschen Datenschutzrecht die Auffassung, dass ein DSGVO-Verstoß ohne spürbare, nachweisbare wirtschaftliche Schäden folgenlos bleibt. Diese Praxis wurde durch die Rechtsprechung des Europäischen Gerichtshofs (EuGH) und des Bundesgerichtshofs (BGH – Urteil vom 18.11.2024 – VI ZR 10/24) grundlegend zerschlagen:

- **Der Grundsatz:** Bereits der **bloße Kontrollverlust** über die eigenen personenbezogenen Daten begründet einen ersatzfähigen immateriellen Schaden nach Art. 82 Abs. 1 DSGVO.
- **Die Auswirkung auf Cloud-KI:** Werden Sprachdaten von Mitarbeitern, Betriebsberichte mit Personennamen oder Kundeninformationen an Server US-amerikanischer Hyperscaler übermittelt, verliert der Betroffene die Kontrolle darüber, wer Zugriff auf diese Daten hat (beispielsweise US-Nachrichtendienste über den FISA 702 / CLOUD Act).
- **Das Risiko:** Betroffene Mitarbeiter oder Kunden benötigen keinen Nachweis eines finanziellen Schadens. Die bloße Ungewissheit und der Kontrollverlust reichen aus, um standardisierte Schadensersatz-

ansprüche (typischerweise 500 bis 2.500 Euro pro Person) geltend zu machen. Bei Belegschaften von 50 bis 500 Mitarbeitern summiert sich dies schnell zu existenzgefährdenden Summen.

Falle C: Der Kollaps der Auftragsverarbeitung (Art. 28 vs. Art. 26 DSGVO)

Nahezu jeder US-Cloud-KI-Anbieter (wie Microsoft für Azure OpenAI oder OpenAI Enterprise) bietet standardmäßig Auftragsverarbeitungsverträge (Data Processing Addenda / AVV nach Art. 28 DSGVO) an und versichert werbewirksam: *„Wir verarbeiten Ihre Daten DSGVO-konform und trainieren unsere Basismodelle nicht mit Ihren Daten."*

In der betrieblichen und gerichtlichen Praxis erweisen sich diese Dokumente häufig als juristischer Papiertiger:

1. **Das Telemetrie- und Diagnose-Schlupfloch (Der aktuelle DSK-Befund):** Selbst wenn der reine Wortlaut des Prompts nicht zum Modelltraining genutzt wird, protokollieren Cloud-Plattformen zwingend umfangreiche System-, Diagnose- und Telemetriedaten (IP-Adressen, Zeitstempel, Stimmmetadaten, Token-Latenzen). Ein Auftragsverarbeiter nach Art. 28 Abs. 3 lit. a DSGVO darf Daten **ausschließlich auf dokumentierte Weisung** des Kunden nutzen. Sobald der Dienstleister Daten für eigene Produkt- oder Sicherheitsanalysen verwendet, erlischt das AVV-Privileg augenblicklich. Es liegt entweder eine unzulässige Drittübermittlung vor oder das Verhältnis mutiert zu einer **Gemeinsamen Verantwortlichkeit (Joint Controllership nach Art. 26 DSGVO)** – mit gesamtschuldnerischer Haftung des Mittelständlers.
2. **US-Bundesrecht sticht EU-Verträge (FISA 702 & US CLOUD Act):** Auch wenn US-Anbieter mit der *EU Data Boundary* und Rechenzentrumsstandorten in Frankfurt am Main werben: Als US-Konzerne unterliegen sie dem US-Bundesrecht. Gemäß Section 702 des Foreign Intelligence Surveillance Act (FISA) und dem Clarifying Lawful Overseas Use of Data Act (CLOUD Act) sind US-Unternehmen gesetzlich verpflichtet, US-Sicherheitsbehörden Zugriff auf gespeicherte Daten zu gewähren. Ein privatrechtlicher AVV kann amerikanisches Bundesrecht niemals suspendieren.
3. **Der Haftungstrick:** In den AGB der Hyperscaler ist die Haftung in der Regel auf die Höhe der im letzten Jahr gezahlten Softwaregebühren gedeckelt. Drohen Aufsichtsbehörden-Bußgelder nach Art. 83 DSGVO (bis zu 20 Millionen Euro) oder machen Mitarbeiter Schadensersatz wegen BGH-Kontrollverlusts (Art. 82 DSGVO) geltend, verklagen sie nicht den Software-Konzern in Übersee, sondern die lokale Geschäftsführung.
4. **Der AVV bindet den Betriebsrat nicht (§ 87 Abs. 1 Nr. 6 BetrVG):** Ein AVV regelt rein das Verhältnis zwischen dem Kundenunternehmen und dem Softwarehersteller. Gegenüber der Belegschaft und dem Betriebsrat besitzt er keinerlei Legitimationswirkung.

1.3 Der Betriebsrat als Gatekeeper (§ 87 Abs. 1 Nr. 6 BetrVG)

Nach § 87 Abs. 1 Nr. 6 BetrVG hat der Betriebsrat ein zwingendes Mitbestimmungsrecht bei der Einführung technischer Einrichtungen, die objektiv zur Überwachung von Verhalten oder Leistung geeignet sind.

Kriterium	Externes AI-Call-Center (Kunde ruft an)	Interne Sprach-KI (Mitarbeiter vor Ort)
Freiwilligkeit	**Gegeben.** Anrufer wird per Bandansage gewarnt und willigt konkludent ein.	**Unmöglich.** Im Beschäftigungsverhältnis gilt das Über-/Unterordnungsverhältnis (§ 26 BDSG).
AI Act Status	**Reine Transparenz (Art. 50).** Support-Bot ist kein Hochrisiko-System.	**Gefahr von Hochrisiko (Anhang III).** Dokumentation von Arbeitsleistung droht als HR-KI.
Betriebsrat	Nicht betroffen (keine internen Mitarbeiter).	**Zwingendes Veto nach § 87 BetrVG.** Stimme ist biometrisch überwachungsfähig.

BGH-Haftung	Einwilligung deckt Risiko ab.	**Akut:** Mitarbeiter wird zur Nutzung verpflichtet; Kontrollverlust greift.

1.4 Risikoklassifizierung nach dem EU AI Act im Betrieb

Die verunsicherte Frage vieler Geschäftsführer: *„Müssen wir unsere lokale Sprach-KI jetzt teuer einer Konformitätsbewertung unterziehen und mit dem CE-Zeichen zertifizieren?“*
Die Antwort lautet: **Eindeutig Nein.**

1. **Keine biometrische Identifizierung (Art. 4 Nr. 14 DSGVO):** Das bloße Erfassen von Schallwellen zur reinen Texttranskription (*Speech-to-Text*) stellt keine biometrische Identifizierung dar. Es wird kein Stimmprofil erzeugt. Der Audio-Buffer im flüchtigen RAM wird sofort atomar mit Nullen überschrieben (*Secure Memory Wipe*).
2. **Reine Narrow-AI ohne Hochrisiko-Klassifizierung:** Die CE-Kennzeichnungspflicht gilt ausschließlich für Hochrisiko-Systeme (Anhang I und III) sowie GPAI mit Systemrisiken. Lokale Sprach-zu-JSON-Modelle sind **Minimal-Risk-Systeme** ohne CE-Zertifizierungspflicht.

1.5 Der architektonische Befreiungsschlag: 100 % Lokale Edge-Inferenz

Die beste juristische Verteidigung ist keine 80-seitige Vereinbarung mit Haftungsausschlüssen, sondern eine **Architektur, die das Entstehen von Risiken physikalisch ausschließt** (*Privacy by Design & by Default*, Art. 25 DSGVO).

Kriterium	Konventionelle Cloud-KI (OpenAI, MS Copilot)	Synapse Delić Edge-Architektur
Datenabfluss	Daten fließen in fremde Rechenzentren.	**Null Bytes verlassen das Gerät.**
Drittstaatentransfer	Hochriskant (FISA 702 / CLOUD Act).	**Entfällt vollständig.** Kein Datentransfer.
BGH-Kontrollverlust	Voll eröffnet (Art. 82 DSGVO).	**Ausgeschlossen.** Datenhoheit bleibt vor Ort.
Betriebsrat	Extrem schwierig / Blockadegefahr.	**Reibungslos.** Keine dauerhafte Audiospeicherung.
Ausfallsicherheit	Anfällig für Funklöcher und Timeouts.	**100 % autark.** Rechnet im tiefsten Keller.

1.6 Praxis-Checkliste vor dem KI-Rollout

1. **Rollenbestimmung:** Werden Modelle für personal- oder prozesskritische Bewertungen eingesetzt?
2. **Datenfluss-Prüfung:** Wohin fließen Prompts und Audios? (Gefahr von Art. 26 Joint Controllership).
3. **Betriebsrats-Check:** Können Rückschlüsse auf Leistung gezogen werden? Rohe Audiodaten niemals speichern.
4. **BGH-Haftungsschutz:** Bei Mitarbeiter- und Betriebsdaten striktes Verbot von Cloud-APIs.
5. **Souveränität:** Auf Open-Weight-Modelle (Llama, Whisper, Mistral) auf lokaler Hardware setzen.

Takeaway: Recht & Compliance

- **Duales Rechtsregime beachten:** Der EU AI Act heilt Datenschutzverstöße nicht. Die DSGVO bleibt nach Art. 2 Abs. 7 uneingeschränkt verbindlich – ohne eigenständige DSGVO-Rechtsgrundlage nützt auch das beste CE-Zertifikat nichts.
- **Haftungsfalle Zweckänderung („Frisiertes Mofa"):** Wer Standard-Cloud-Modelle für Personal- oder Prozessentscheidungen konfiguriert, wird nach Art. 25 AI Act automatisch zum voll haftenden Hersteller (Provider) mit allen Pflichten nach Anhang III.
- **BGH-Schadensersatz ausschließen:** Der Bundesgerichtshof urteilt (Az. VI ZR 10/24): Schon der bloße Datenkontrollverlust durch Übertragung an US-Clouds begründet immateriellen Schadensersatz (Art. 82 DSGVO).
- **Betriebsrat als Gatekeeper mitnehmen:** Weil Sprache biometrisch wirkt (§ 87 Abs. 1 Nr. 6 BetrVG), scheitern Cloud-Voicebots. Lokale Edge-Modelle mit flüchtigem RAM-Wipe erfordern keine Tonaufzeichnung und sind betriebsratsfest.
- **Keine CE-Zertifizierungspflicht:** Lokale Narrow-AI (Speech-to-Text & JSON-Strukturierung) fällt unter Minimal Risk nach dem AI Act. Es sind weder CE-Kennzeichnung noch teure Audits erforderlich.

1.7 Quellen & Rechtsprechungs-Referenzen

- **DSK Strategieklausur Speyer (07.09.2026):** Konsequente Durchsetzung bei neuen Technologien und biometrischer Sensorik.
- **DSK Beschluss zur KI-Verantwortung (23.04.2026):** Volle Betreiberhaftung; Warnung vor Risiko-Abwälzung auf KMU.
- **DSK Digital Fitness Check (17.03.2026):** Reforminitiative zur KMU-Entlastung bei gleichzeitiger Plattform-Schärfung.
- **DSK Orientierungshilfe RAG (Oktober 2025):** Risikominimierung durch lokale Datenhaltung und On-Premise/Edge-Modelle.
- **DSK Orientierungshilfe TOM (Juni 2025):** Vorgaben für flüchtige Speicherpuffer (RAM) und atomare Memory Wipes.
- **DSK Orientierungshilfe KI (06.05.2024):** Grenzen von Standard-AVVs und Rechtsgrundlagen.
- **BGH, Urteil vom 18.11.2024 – VI ZR 10/24:** Kontrollverlust begründet immateriellen Schaden nach Art. 82 DSGVO.
- **BAG, Beschluss vom 13.12.2016 – 1 ABR 7/15:** Zwingende Mitbestimmung nach § 87 Abs. 1 Nr. 6 BetrVG.

SYNAPSE DELIĆ · PRAXISBUCH

Kapitel 2: Vom Rechenzentrum ins Funkloch: Das technische How-To

> **Szene vor Ort: Das Laderädchen im zweiten Untergeschoss**
>
> *Dienstagmorgen, 07:15 Uhr. Polier Dragan steht im 2. Untergeschoss eines Klinik-Rohbaus. Aus einer Deckenfuge tropft Schichtenwasser direkt auf die frisch installierten Steigtrassen – akute Behinderungsanzeige nach § 6 Abs. 1 VOB/B erforderlich. Er zückt das Smartphone, öffnet das cloudbasierte Bautagebuch des Auftraggebers und spricht die Mängelrüge ein. Auf dem Display erscheint das Ladesymbol: „Übertrage Audiodaten… Verbindung wird gesucht… Fehler 504 Gateway Timeout." Wegen des massiven Stahlbetons herrscht Funkstille. Entnervt steckt Dragan das Smartphone weg, kritzelt zwei unvollständige Notizen auf einen staubigen Zementsack und vergisst abends im Feierabend die Hälfte. Vier Monate später weist der Bauherr eine Nachforderung über 42.000 Euro ab, weil der kausale Behinderungsnachweis lückenhaft war. Mit lokaler Edge-KI hätte Dragans Spracheingabe in 280 Millisekunden direkt im lokalen Speicher zu einem gerichtsfesten JSON-Protokoll geführt – ohne einen einzigen Balken Netz.*

Sobald digitale Prozesse in die reale Wertschöpfung vordringen – auf Baustellen, in Keller, landwirtschaftliche Betriebe oder Transport-LKWs –, bricht die Cloud-Doktrin an vier physikalischen Realitäten zusammen:

- **Funkloch:** Im 2. Untergeschoss gibt es weder 5G noch WLAN.
- **Latenz:** 2 bis 4 Sekunden Latenz für transatlantische API-Aufrufe frustrieren Mitarbeiter im Gehen.
- **Energie- und Rechenzentrums-Verschwendung:** 70B-Modelle auf Dual-A100-Grafikkarten besitzen Weltwissen über Shakespeare, sind für einfache Bau- oder Lieferscheinmemos aber grotesk überdimensioniert.
- **Kostenfalle:** Variable Token-Gebühren erzeugen unkalkulierbare Kostenkurven.

Die Lösung: **Sovereign Edge AI** – kompakte Fach-Modelle, die direkt auf mobilen Chips ohne Netzanbindung in Echtzeit rechnen.

2.1 Die 4-Stufen-Schrumpf-Pipeline: Von 140 GB auf 500 MB

1. **Knowledge Distillation (Lehrer-Schüler-Verfahren):** Ein 70B-Modell im Cloud-Cluster trainiert ein kompaktes 1.5B–3B Schülermodell gezielt auf Fachjargon (z. B. VOB, Mängelrügen, Pflegedokumentation). Der Schüler erlernt die logische Struktur und Terminologie, vergisst jedoch unnötiges Allgemeinwissen.
2. **Vokabular-Pruning:** Das Tokenizer-Embedding wird von 128.000 Token auf 32.000 relevante deutsche und fachtechnische Token reduziert. Das spart über 600 Megabyte Arbeitsspeicher.
3. **INT4-Quantisierung (GGUF / AWQ):** Gewichte werden von 16-Bit-Kommazahlen auf 4-Bit-Ganzzahlen gerundet (**75 Prozent Speicherersparnis**). Der Genauigkeitsverlust im Fachbereich liegt unter 1,5 Prozent.
4. **Hardware-Kompilierung & Zero-Copy `mmap`:** Das Modell läuft auf der Neural Engine (NPU) oder GPU via Apple CoreML / ONNX Runtime / llama.cpp und startet via Memory-Mapping in 50 Millisekunden ohne Ladeverzögerung.

2.2 Die Zero-UI-Systemarchitektur: Voice-to-JSON Pipeline

- **VAD (Voice Activity Detection):** Filtert Baustellen- und Windgeräusche in unter 10 Millisekunden heraus.
- **Streaming STT (Whisper INT4):** Transkribiert Sprache 5x schneller als Echtzeit lokal auf der NPU.

- **Strukturierung mit GBNF-Grammar:** Das Small Language Model wird durch formale Grammatiken mathematisch gezwungen, **ausschließlich syntaktisch valides JSON** zu generieren. Halluzinationen und fehlerhafte Felder sind architektonisch unmöglich.
- **Ephemerer RAM-Lifecycle:** Audiodaten existieren nur für Millisekunden im flüchtigen RAM und werden nach der JSON-Extraktion sofort atomar mit Nullen überschrieben (**Secure Wipe**).

2.3 Hardware-Dimensionierung für den Mittelstand

Einsatzklasse	Hardware-Beispiele	RAM-Bedarf	Typische Latenz
Klasse A: Mobil (Bauleiter, Fahrer, Monteure)	iPhone 15/16 Pro, Galaxy S24, Pixel 8/9, Zebra-Industrie-Tablets	450 MB – 1,8 GB	**< 400 ms** (Flugmodus autark)
Klasse B: Industrie-Edge (Schaltschrank, Werkshalle)	Raspberry Pi 5 mit Hailo-8L, Lüfterloser Industrie-PC (Intel N100)	1,2 GB – 2,5 GB	**< 600 ms** (12V Dauerbetrieb)
Klasse C: Office-Gateway (Zentrale, Dispatcher, 50 Nutzer)	Apple Mac Studio (M2/M4 Max), Mini-Server mit 1x Nvidia RTX 4060 Ti	8 GB – 16 GB	**< 250 ms** (Zentral im LAN)

Takeaway: Edge-Technologie & Zero-UI

- **Physik vor Cloud-Versprechen:** Auf Baustellen, in Werkshallen und Tiefgaragen herrscht das Funkloch. Sprach-KI für den operativen Betrieb muss 100 % autark und ohne Internetverbindung in unter 300 ms rechnen.
- **Vier-Stufen-Schrumpf-Pipeline:** Durch Knowledge Distillation, Vokabular-Pruning und INT4-Quantisierung schrumpfen Modelle von 140 GB auf unter 2 GB – bei vollem Fachwortschatz und weniger als 1,5 % Genauigkeitsverlust.
- **Deterministische Präzision (GBNF):** Formale Grammatiken zwingen das Sprachmodell mathematisch in ein valides JSON-Schema. Halluzinationen und ungültige ERP-Eingaben sind architektonisch ausgeschlossen.
- **Volle Hardware-Kompatibilität:** Keine teuren Cloud-Server nötig. Edge-Inferenz läuft direkt auf handelsüblichen iPhones, Android-Smartphones oder lüfterlosen 15-Watt-Industrie-PCs.

SYNAPSE DELIĆ · PRAXISBUCH

Kapitel 3: Das Synapse-Delić-Vorgehensmodell

> **Der Beratungs-Klassiker: 80 bunte Folien und null Funktion**
>
> *Ein mittelständischer Generalunternehmer beauftragte eine renommierte IT-Unternehmensberatung, um die Baudokumentation in LPH 8 zu digitalisieren. Nach vier Monaten intensiver „Scoping-Workshops“ und 65.000 Euro Beraterhonorar präsentierten drei Consultants im Besprechungsraum eine 90-seitige PowerPoint-Präsentation mit komplexen Reifegradmodellen und einem 14-monatigen ERP-Transformationsplan. Als der erfahrene Oberbauleiter fragte: „Können meine Jungs das ab Montag auf der Baustelle ausprobieren?“, herrschte betretenes Schweigen: Es gab noch keine einzige Zeile Code.*
>
> *Synapse Delić bricht radikal mit dieser Kultur: Keine Beraterfolien, sondern mit Sicherheitsschuhen raus auf die Baustelle. 45 Minuten echtes Shadowing im Dreck, und 14 Tage später hält die Bauleitung einen funktionierenden, offline-fähigen Prototyp auf echter Hardware in der Hand.*

Klassische IT-Berater verlangen 50.000 bis 100.000 Euro für bunte Strategie-Folien, können aber keine Zeile lauffähigen Code schreiben. Klassische Systemhäuser bauen überfrachtete Menüs, die nach monatelangen Verzögerungen zur ungenutzten Shelfware werden.
Synapse Delić folgt den Prinzipien des **Mom-Tests (Rob Fitzpatrick)** und der **No-oriented Discovery (Chris Voss)**: Wir gehen dorthin, wo im Dreck gearbeitet wird. Wir beobachten den realen Arbeitsfluss und identifizieren die echten Medienbrüche, bevor eine einzige Zeile Code geschrieben wird.

3.1 Der 14-Tage High-Speed Zero-UI MVP

Auf Basis unserer horizontalen Core-Engines (`Voice-2-x` und `Interface-2-x`) liefern wir innerhalb von **14 Werktagen** einen funktionierenden, offline-fähigen Prototyp auf echter Hardware:
- Keine hypothetischen Mockups.
- Ein lauffähiges System, das der Polier, Fahrer oder Monteur sofort im Arbeitsalltag testen kann.

3.2 Software-Ownership statt SaaS-Abofalle

Mittelständler wollen keine endlosen monatlichen Lizenzketten pro Mitarbeiter bezahlen, bei denen jede kleine Änderung zur teuren Change-Request-Falle wird:
- Synapse Delić liefert **Software-Ownership (Buyout / One-Off)**.
- Der Kunde behält die volle Hoheit über Modelle, Daten und Code.
- Modulare Anpassungen und Wartung erfolgen agil in Tagen statt Monaten.

> **Takeaway: Strategie & Software-Ownership**
>
> - **Mom-Test Discovery vor Code:** Keine theoretischen 80-Seiten-Folien von teuren Beratern. Wir gehen in den Dreck der Baustelle, beobachten reale Handgriffe und decken die echten Medienbrüche auf.
> - **14-Tage High-Speed Zero-UI MVP:** Schneller, lauffähiger Prototyp auf echter Hardware in zwei Wochen statt sechsmonatiger Feature-Factories, die zur ungenutzten Shelfware verkommen.
> - **Software-Ownership statt SaaS-Abofalle:** Einmalige Entwicklung mit vollständiger Eigentümerschaft an Code, Modellen und Daten. Keine steigenden Lizenzkosten pro Kopf und keine Change-Request-Erpressung.

SYNAPSE DELIĆ · PRAXISBUCH

Kapitel 4: Abschluss & Ausblick: Der Weg zur betrieblichen KI-Souveränität

> **Ausblick: Wer heute die Architektur richtig wählt, sichert sich die nächsten zehn Jahre**
>
> *Die entscheidende strategische Frage für den deutschen Mittelstand lautet im Jahr 2026 nicht mehr, ob Künstliche Intelligenz Einzug in den operativen Betrieb hält, sondern auf welchem rechtlichen und physikalischen Fundament sie ruht. Wer sich unkritisch an US-Hyperscaler bindet, zahlt dreifach: durch unkalkulierbare variable Token-Kosten, durch das permanente BGH-Schadensersatzrisiko und durch die Blockade der eigenen Belegschaft. Wer hingegen auf sovereign Edge-Architekturen setzt, schafft ein echtes Unternehmens-Asset, das dauerhaft unabhängig macht und den operativen Vorsprung sichert.*

Erfolgreiche Digitalisierung vor Ort folgt keinen Software-Hypes, sondern handfesten betrieblichen Naturgesetzen. Geschäftsführer und Vorstände sollten jedes KI-Vorhaben an drei Prüfsteinen messen:

1. **Regel 1: Datenhoheit ist unteilbar (Zero-Cloud im operativen Kern)** Mitarbeiterstimmen, Baustellenprotokolle, Mängelrügen und betriebliche Kalkulationen dürfen niemals als Trainingsfutter oder Telemetriedaten auf US-Server abfließen. Lokale Inferenz mit ephemerem RAM-Lifecycle (atomares Überschreiben nach der JSON-Strukturierung) schützt zu 100 Prozent vor BGH-Schadensersatzklagen (Art. 82 DSGVO) und sichert die reibungslose Freigabe durch den Betriebsrat (§ 87 Abs. 1 Nr. 6 BetrVG).
2. **Regel 2: Zero-UI schlägt jeden Menü-Dschungel** Wenn ein Polier, Handwerker, Monteur oder Fahrer mehr als zwei Klicks benötigt, um einen Sachverhalt zu dokumentieren, wird die Software auf der Baustelle sterben. Software muss sich der menschlichen Sprache anpassen – nicht der Mensch einer komplizierten ERP-Maske.
3. **Regel 3: Software-Ownership vor SaaS-Knechtschaft** Mieten Sie keine Standard-Software für Ihre geschäftskritischen Kernprozesse. Bestehen Sie auf dem vollen Eigentum an Quellcode, Modellen und Gewichten. Nur mit echter Software-Ownership bleiben Sie unabhängig von Preissteigerungen und setzen individuelle Change Requests in Tagen statt in Monaten um.

4.1 Ausblick 2026/2027: Die Evolution der Edge-Intelligenz

Während viele Unternehmen noch mit langsamen Cloud-Chatbots experimentieren, rollt bereits die nächste Welle der Edge-Technologie an:

- **Hardware-Standardisierung:** Nahezu jedes moderne Smartphone und Industrie-Tablet besitzt heute dedizierte neuronale Beschleuniger (NPU / Neural Engine). Inferenz-Latenzen von unter 100 Millisekunden werden zum Alltag – bei minimalem Akkuverbrauch und 100 % Offline-Verfügbarkeit.
- **On-Device Kontextgedächtnis (Local RAG):** Kompakte Sprachmodelle greifen künftig direkt auf dem Endgerät auf lokale Projektpläne, VOB-Leistungsverzeichnisse oder Maschinenhandbücher zu, ohne ein einziges Byte ins Netz zu senden.
- **Rezeptive Mehrsprachigkeit:** Ausländische Fachkräfte sprechen in ihrer Muttersprache (z. B. Serbisch, Polnisch oder Türkisch) frei ins Gerät – die lokale KI erfasst den Sachverhalt und erzeugt in Echtzeit ein fehlerfreies, VOB-konformes Fachprotokoll auf Deutsch.

Takeaway: Der 14-Tage-Aktionsplan

1. **Schritt 1 (Tag 1–3) – Mom-Test Discovery:** Keine 80 Folien. 45 Minuten stilles Über-die-Schulter-Schauen vor Ort. Den größten realen Medienbruch zwischen Frontline und Büro identifizieren.
2. **Schritt 2 (Tag 4–14) – Zero-UI MVP:** Umsetzung eines kompakten, offline-fähigen Prototyps auf Basis spezialisierter Open-Weight-Modelle (Whisper / SLMs) direkt auf den Endgeräten der Mitarbeiter.
3. **Schritt 3 (Tag 15) – Praxistest im Funkloch:** Das System wird unter realen Bedingungen im Keller, im Tunnel oder auf der Baustelle getestet. Erst wenn die Frontline begeistert ist, erfolgt der Rollout.
4. **Schritt 4 – Dauerhafte Souveränität:** Übernahme der Lösung in das eigene Betriebsvermögen (Software-Ownership). Volle Daten- und Codehoheit ohne laufende Pro-Kopf-Abo-Fallen.

Direkter Download & Unverbindliches Sparring

Scannen Sie diesen QR-Code mit der Smartphone-Kamera, um das PDF digital abzurufen oder direkt ein vertrauliches 15-Minuten-Strategie-Sparring mit Emil Delić zu vereinbaren.

https://synapse-delic.de/whitepaper

Synapse Delić UG (haftungsbeschränkt) · 38518 Gifhorn
E-Mail: mail@synapse-delic.de · Tel: +49 1573 1593419

© 2026 Synapse Delić UG (haftungsbeschränkt). Alle Rechte vorbehalten.