Edge-Privacy und Transparenz - Der Weg zu einem verantwortungsvollen KI-Tagesbegleiter

Alexander Lutz
FOM Hochschule für Oekonomie & Management

Summary

Körpernahe KI-Audioassistenten markieren den Übergang von punktueller Sprachinteraktion zu permanenten, kontextsensitiven Begleitern. Daraus erwachsen Risiken: intransparente Aufzeichnungen im öffentlichen Raum, unklare Einwilligung Dritter, schwer auditierbare Datenflüsse und unsichere Update-Kanäle. Der Vortrag präsentiert einen europäischen Gegenentwurf: einen prototypischen KI-Tagesbegleiter („NeuroMate AI“) als Governance-Blueprint. Technisch setzt das System auf Edge-Processing (Whisper-Inference lokal), Datensparsamkeit und Offline-First; organisatorisch auf Privacy-/Security-by-Design, reproduzierbare Builds (Docker), signierte und versionierte Updates mit schneller Rückkehr auf die letzte stabile Version.

Zentral sind Transparenz- und Einwilligungsmechanismen: ein hardware-gebundener Aufnahme-Indikator (sicht-/hörbar), Wake-Word-Gating, sofortiges Opt-Out für Umstehende sowie standardisierte Statussignale nach Vorbild sichtbarer Kennzeichnungen in anderen Domänen. Ergänzt wird dies durch Auditierbarkeit (lokale Protokolle, klare Datenhaltungsgrenzen), Rollen- und Profilmanagement, sowie Policy-Hooks für organisatorische Richtlinien.

Wir diskutieren Designentscheidungen, Trade-Offs (Akkulaufzeit vs. On-Device-Schutz), erste Messergebnisse (Latenz, Erkennungsraten, Robustheit in Alltagsumgebungen) sowie einen Vorschlag für pragmatische Mindeststandards: sichtbare Aufnahme-Signale, Default-Mute im öffentlichen Raum, dokumentierte Datenpfade, verpflichtende Update-Transparenz, und ein Label für vertrauenswürdige Wearable-KI.

Der Beitrag beleuchtet Compliance als Innovationsmotor: wer Verantwortung früh in Architektur und UX verankert, baut schneller Vertrauen auf, reduziert Haftungsrisiken und schafft einen Wettbewerbsvorteil für europäische Anbieter. Die Session schließt mit einer Live-Demonstration des Prototyps und konkreten Empfehlungen für Bildung, Beschaffung und Standardisierung.

Keywords:

Compliance, Edge-Processing, Edge-AI, Datenschutz-By-Design, Wake-Word

1. Einleitung

Der Übergang von reaktiven Smart Speakern zu proaktiven, permanenten KI-Wearables markiert einen Paradigmenwechsel in der Mensch-Maschine-Interaktion. Während stationäre Systeme wie Amazon Echo oder Google Home auf punktuelle Sprachbefehle in definierten Räumen warten und aufgrund ihrer Always-on-Mikrofone bereits erhebliche Datenschutzfragen aufwerfen (Malkin et al., 2019; Hernández Acosta & Reinhardt, 2022), begleiten körpernahe KI-Assistenten ihre Nutzer kontinuierlich durch den Alltag. Diese Entwicklung knüpft an Grundideen des Context-Aware Computing an, bei dem Systeme ihr Verhalten an situative Informationen über Nutzer, Umgebung und Interaktionskontext anpassen (Dey et al., 2001). Aktuelle Arbeiten zu proaktiven Wearable-AI-Systemen zeigen, dass solche Geräte nicht mehr nur auf explizite Kommandos reagieren, sondern kontextabhängige Hinweise, Erinnerungen oder Interventionen anbieten können (Fang et al., 2025). Diese "Always-on"-Systeme versprechen eine nahtlose, kontextsensitive Unterstützung, indem sie Gespräche, Meetings und Umgebungsgeräusche erfassen, transkribieren und mithilfe von Large Language Models analysieren (Takhshid, 2024).

Diese technologische Evolution kollidiert jedoch massiv mit europäischen Datenschutzgrundsätzen, insbesondere der Datenschutz-Grundverordnung (DSGVO) und dem neuen EU AI Act (European Union, 2016, 2024). Aktuelle kommerzielle Lösungen wie der „AI Pin" oder das „Pocket" AI-Wearable zur Gesprächsaufzeichnung basieren vielfach auf Cloud- beziehungsweise Drittanbieter-Inferenz. Beim Humane „AI Pin" wurde der Zugriff auf Modelle wie OpenAI ChatGPT und Google Gemini beschrieben; bei Limitless wird in der Privacy Policy ausdrücklich angegeben, dass Audioaufnahmen oder Transkripte an Drittanbieter-KI-Dienste zur Transkription und Zusammenfassung übermittelt werden können (Limitless AI, 2025; Chokkattu, 2024). Dies führt zu gravierenden Datenschutz- und Sicherheitsproblemen: Intransparente Datenflüsse, die Speicherung sensibler Informationen in Drittstaaten, technische Angriffsmöglichkeiten auf Sprachassistenzsysteme sowie policy-bezogene Privacy Breaches wurden in der Forschung zu Voice-Assistant-Anwendungen bereits umfassend beschrieben (Li et al., 2023). Besonders kritisch ist dabei die Verletzung der sogenannten „Bystander Privacy", also des Schutzes unbeteiligter Dritter, deren Gespräche ohne explizite Einwilligung erfasst werden können. Takhshid (2024) beschreibt genau dieses Problem für Wearable-AI-Geräte mit audiovisuellen Erfassungsfunktionen und argumentiert für „privacy frictions" als sichtbare, greifbare Hinweise und Einwilligungsmechanismen für Umstehende.

Gleichzeitig zeigen aktuelle Forschungsarbeiten, dass die technologischen Grundlagen für eine datenschutzfreundliche Alternative bereits existieren. Die rasante Entwicklung im

Bereich Edge AI ermöglicht es zunehmend, komplexe Machine-Learning-Modelle direkt auf ressourcenbeschränkten Geräten auszuführen (Somvanshi, 2025). Fortschritte bei Tiny Foundation Models, effizienter On-Device Automatic Speech Recognition (ASR) und Offline-First-Architekturen belegen, dass eine lokale Verarbeitung von Sprachdaten ohne Cloud-Abhängigkeit machbar ist (Benazir & Lin, 2025; Dey, 2024). So demonstrieren aktuelle Studien, dass leichtgewichtige Transformer-Modelle und quantisierte ASR-Systeme auf Edge-Geräten eine Genauigkeit erreichen können, die serverseitigen Lösungen nahekommt, während sie gleichzeitig Latenz und Energieverbrauch optimieren (Gondi, 2021; Torkamani, 2025).

Das Forschungsziel dieses Beitrags ist die Vorstellung eines prototypischen KI-Tagesbegleiters - des “NeuroMate AI” - als Referenzarchitektur für Privacy-by-Design in Edge-AI-Systemen. Der NeuroMate AI demonstriert, dass eine sinnvolle, DSGVO-konforme Variante möglich ist: Offline-First Architektur, lokale KI-Inferenz, Nutzung starker Open-Source-Modelle und ein rollenbasiertes Rechtemanagement. Anstatt Bequemlichkeit auf Kosten der Privatsphäre zu erkaufen, zeigt dieser Prototyp, wie technologische Innovation und europäische Werte in einem verantwortungsvollen Wearable vereint werden können.

2. Regulatorischer und ethischer Kontext

Die Entwicklung und der Einsatz von KI-Wearables bewegen sich in einem komplexen Spannungsfeld zwischen technologischer Machbarkeit und strengen regulatorischen Vorgaben. Der europäische Rechtsrahmen, insbesondere definiert durch die DSGVO und der EU AI Act, setzt klare Grenzen für die allgegenwärtige Datenerfassung, bietet aber gleichzeitig einen Gestaltungsrahmen für vertrauenswürdige Innovationen.

Ein zentraler Aspekt ist das Prinzip der Datensparsamkeit (Art. 5 DSGVO) und das Gebot des Privacy-by-Design (Art. 25 DSGVO) (Cavoukian, 2009). Diese Prinzipien fordern, dass Datenschutz nicht nachträglich implementiert, sondern von Beginn an in die Systemarchitektur integriert wird. Cloud-basierte Wearables, die kontinuierlich Audiodaten an externe Server streamen, verstoßen strukturell gegen das Prinzip der Datensparsamkeit, da oft mehr Daten erhoben und übertragen werden, als für den eigentlichen Zweck zwingend erforderlich wäre. Dies ist insbesondere im Rahmen der Transparenzpflichten (Art. 50 des EU AI Act) für KI-Systeme. Im Gegensatz dazu setzt der NeuroMate AI auf Edge-Processing, bei dem die Daten das Gerät nicht verlassen, was das Risiko von Datenabflüssen (Data Leaks) drastisch minimiert und die DSGVO-Konformität by Design sicherstellt.

Eine besondere ethische und rechtliche Herausforderung bei Wearables ist die “Bystander Privacy” (Takhshid, 2024), welche die Einwilligung Dritter betrifft. Im öffentlichen Raum oder in (Online-) Meetings erfassen Mikrofone unweigerlich die Stimmen von Personen, welche dem System weder zugestimmt haben noch sich dessen Präsenz bewusst sind. Die aktuelle Gesetzgebung stößt hier an ihre Grenzen, da die Einholung einer informierten Einwilligung (Consent) von flüchtigen Passanten oder Gesprächspartnern in Echtzeit kaum praktikabel ist. Kommerzielle Anbieter wälzen diese Verantwortung oft in ihren Nutzungsbedingungen auf den Träger des Geräts ab. Der NeuroMate AI adressiert dieses Problem durch das Konzept der “Tangible Privacy” (Koelle et al., 2020): Hardware-gebundene, sicht- und optional hörbare Aufnahme-Indikatoren sowie ein sofortiges Opt-Out für Umstehende schaffen Transparenz und reduzieren die sogenannten “Privacy Frictions” (Takhshid, 2024).

Zusätzlich verschärft der neue EU AI Act die Anforderungen an KI-Systeme. Artikel 50 des AI Acts formuliert explizite Transparenzpflichten für KI-Systeme, die mit Menschen interagieren oder biometrische Daten verarbeiten (Europäische Union, 2024). Nutzer und Umstehende müssen darüber informiert werden, dass sie mit einer KI interagieren oder von ihr erfasst werden. Ein System, das heimlich im Hintergrund agiert, ist unter diesem Regime nicht zulässig.

Die Einhaltung dieser strengen Vorgaben wird oft als Innovationsbremse wahrgenommen. Dieser Beitrag argumentiert jedoch gegenteilig: Compliance ist ein entscheidender Wettbewerbsvorteil (Musch et al., 2024). In einem Markt, der zunehmend von Misstrauen gegenüber intransparenten Datenpraktiken geprägt ist (wie Analysen der Datenschutzrichtlinien führender Wearable-Hersteller zeigen; Doherty et al., 2025), schafft Privacy-by-Design Vertrauen. Europäische Werte, übersetzt in technische Architektur (Edge-Inference, Offline-First), dienen als Innovationsmotor. Wer Verantwortung frühzeitig in Architektur und User Experience (UX) verankert, reduziert nicht nur Haftungsrisiken, sondern etabliert einen Goldstandard für vertrauenswürdige KI, der durch Zertifizierungen wie ein “Trustworthy Wearable AI Label” (Scharowski et al., 2023) auch kommerziell differenzierend wirken kann.

3. Systemarchitektur des "NeuroMate AI"-Handhelds

In diesem Abschnitt werden sowohl die technische Hardware detailliert beschrieben, das Edge-Processing wird mit einer Begründung für On-Device-Inference zur Vermeidung von Cloud-Abhängigkeiten begründet. Abschließend werden die Umsetzungen von Privacy-by-Design ausgeführt.

3.1 Hardwarearchitektur

Die technische Architektur des NeuroMate-AI-Handhelds folgt dem Grundsatz einer lokal ausführbaren, auditierbaren und datensparsamen KI-Infrastruktur. Im Zentrum steht nicht die permanente Auslagerung sprachlicher Interaktionen in Cloud-Systeme, sondern eine Edge-First-Architektur, bei der Spracherkennung, Kontextverarbeitung, Modellinferenz und Protokollierung weitgehend lokal erfolgen. Diese Entscheidung ist sowohl technisch als auch datenschutzbezogen begründet: Edge Computing reduziert Latenz, Bandbreitenbedarf und Cloud-Abhängigkeit und kann zugleich helfen, Daten näher am Entstehungsort zu kontrollieren (Shi et al., 2016; Satyanarayanan, 2017; Yu et al., 2018). Gerade bei sprachbasierten Systemen ist diese lokale Verarbeitung besonders relevant, da aus Stimme und Sprechweise sensible Informationen wie Alter, Herkunft, Emotionen, Gesundheitszustand oder Persönlichkeitsmerkmale abgeleitet werden können (Kröger et al., 2022).

Die Entwicklungs- und Zielarchitektur ist zweistufig aufgebaut. Einerseits existiert ein leistungsfähiges lokales Host-System, welches als Entwicklungs-, Simulations-, Benchmarking- und Integrationsplattform dient. Dieses System basiert für die locale KI-Inferenz auf einer NVIDIA GeForce RTX 5090. Die RTX 5090 basiert auf der NVIDIA-Blackwell-Architektur, verfügt laut NVIDIA über 21.760 CUDA-Kerne, Tensor Cores der fünften Generation, 32 GB GDDR7-Speicher, ein 512-Bit-Speicherinterface und eine angegebene KI-Leistung von 3.352 AI TOPS. Damit eignet sich das Host-System insbesondere für lokale Modelltests, Quantisierungsvergleiche, Container-Builds, Evaluationsläufe und die Simulation späterer Edge-Deployments (NVIDIA, n.d.). Andererseits ist als mobile Zielplattform ein Banana Pi BPI-M7 vorgesehen. Die Plattform basiert auf dem Rockchip RK3588 mit insgesamt acht CPU-Kernen (vier Cortex-A76-Kerne und vier Cortex-A55-Kernen), einer Mali-G610-MP4-GPU und einer integrierten NPU mit bis zu 6 TOPS INT8. Für das Projekt ist insbesondere die Variante mit bis zu 32 GB LPDDR4/LPDDR4x-RAM, 128 GB eMMC-Speicher und M.2-Key-M-Anbindung relevant. Die M.2-Schnittstelle ist laut Banana-Pi-Dokumentation über PCIe 3.0 mit vier Lanes angebunden und unterstützt NVMe-SSDs. Im Projekt wird diese lokale Speicherarchitektur durch eine Crucial P3 Plus 4 TB M.2 PCIe Gen4 NVMe SSD ergänzt, welche als Modell-, Daten-, Container- und Protokollspeicher dient (Banana Pi Wiki, 2024).

3.2 Lokale Inferenz am Edge und am Hostsystem

Die zusätzliche KI-Beschleunigung auf der Zielplattform wird durch ein Hailo-10H M.2 AI Acceleration Module vorgesehen. Hailo gibt für das Modul 40 TOPS bei INT4 beziehungsweise 20 TOPS bei INT8 an; das Modul ist für Edge-Inferenz konzipiert und soll insbesondere Vision- und generative KI-Anwendungen lokal unterstützen. Relevant ist dabei

nicht nur die nominelle Rechenleistung, sondern auch die Systemintegration über den M.2-Bus. Im Vergleich zu typischen Raspberry-Pi-Erweiterungsszenarien ist die PCIe-3.0-x4-Anbindung des Banana Pi M7 ein wichtiger architektonischer Vorteil, da Host, NVMe-Speicher und Beschleunigermodul mit deutlich höherem Durchsatz miteinander kommunizieren können (Banana Pi Wiki, 2024; Hailo, n.d.). Die praktische Inferenzarchitektur folgt darüber hinaus einem hybriden Modell. Das Host-System mit RTX 5090 dient als lokale „KI-Werkbank“: Dort werden Docker-Container hergestellt, Open Source KI Modelle geladen, Quantisierungsvarianten verglichen, Prompts und Rollenprofile getestet, Evaluationsdaten erzeugt und reproduzierbare Deployments vorbereitet. Die mobile Edge-Hardware übernimmt im späteren Betrieb die echtzeitnahe Interaktion: Wake-Word-Erkennung, Audioaufnahme, lokale Transkription, Intent-Erkennung, Rollen-/Profilauswahl, lokale Retrieval-Prozesse und gegebenenfalls kompakte LLM-Inferenz. Die Trennung von Entwicklungs-Host und Edge-Zielsystem erlaubt es, rechenintensive Optimierungs- und Testprozesse auf der RTX 5090 durchzuführen, ohne die spätere mobile Laufzeitarchitektur unnötig aufzublähen.

3.2.1 Auswahl der lokalen Sprachmodelle (LLMs)

Als lokale Sprachmodellkomponente wird Gemma 4 über Docker und Ollama integriert. Gemma 4 wird von Google DeepMind als Familie offener Modelle beschrieben; sie umfasst unter anderem E2B, E4B, 26B A4B und 31B (Google AI for Developers, 2026). Die kleineren Modelle sind explizit für effiziente On-Device- beziehungsweise Edge-Szenarien gedacht, während die 26B- und 31B-Varianten für leistungsfähige lokale Workstations beziehungsweise Consumer-GPUs geeignet sind. Google beschreibt für Gemma 4 außerdem multimodale Fähigkeiten, größere Kontextfenster und verschiedene Präzisionsstufen einschließlich Quantisierung; für die Inferenz nennt Google beispielsweise Q4_0-Speicherabschätzungen für die einzelnen Modellgrößen.

Für das NeuroMate-System ergibt sich daraus eine sinnvolle Modellstaffelung. Auf dem Host-System mit RTX 5090 können größere Gemma-4-Varianten, etwa 26B A4B oder 31B, lokal getestet und für Workstation-nahe Aufgaben genutzt werden. Auf dem mobilen Edge-Gerät kommen bevorzugt kleinere oder stärker quantisierte Varianten zum Einsatz, insbesondere E2B oder E4B beziehungsweise passende Q4-/Q5-Quantisierungen. Ollama unterstützt die Quantisierung von FP16- beziehungsweise FP32-Modellen über `ollama create --quantize` und nennt unter anderem `q4_K_M`, `q5_K_M`, `q6_K` und `q8_0` als unterstützte Quantisierungsstufen. Dadurch kann je nach Zielhardware zwischen Genauigkeit, Speicherbedarf, Energieverbrauch und Antwortlatenz abgewogen werden. Die wissenschaftliche Begründung für diese Quantisierungsstrategie ergibt sich aus der Forschung zu effizienter neuronaler Inferenz. Quantisierung reduziert Speicherbedarf und Rechenaufwand, kann dadurch lokale Inferenz auf ressourcenbeschränkten Systemen ermöglichen

und ist ein zentrales Verfahren für Edge-AI-Deployments (Gholami et al., 2022). Arbeiten wie GPTQ und QLoRA zeigen zudem, dass niedrigbitige Repräsentationen insbesondere bei großen Sprachmodellen erhebliche Speicher- und Effizienzvorteile erschließen können, ohne die Modellqualität zwangsläufig proportional zu verschlechtern (Dettmers et al., 2023; Frantar et al., 2023). Für NeuroMate bedeutet dies, dass die Auswahl der Quantisierung nicht statisch erfolgen sollte, sondern empirisch anhand von Latenz, Energiebedarf, Antwortqualität, Speicherverbrauch und thermischem Verhalten evaluiert werden muss.

3.2.2 Audioverarbeitung (STT) und lokale semantische Verarbeitung

Die Audioverarbeitung beginnt mit einem Wake-Word-Gating. Das Gerät soll nicht dauerhaft vollständige Sprachdaten transkribieren, sondern zunächst lokal auf ein Aktivierungsereignis reagieren. Erst nach erfolgreicher Aktivierung wird ein begrenztes Audiosignal verarbeitet. Technisch knüpft dieser Ansatz an Keyword-Spotting- beziehungsweise Wake-Word-Detection-Verfahren an, die in sprachbasierten Mensch-Technik-Schnittstellen genutzt werden, um zwischen passiven und aktiven Gerätezuständen zu wechseln. Für eingebettete Systeme sind dabei insbesondere geringe Modellgröße, niedrige Latenz und eine niedrige False-Alarm-Rate relevant (Arık et al., 2017). Zugleich zeigen Untersuchungen zu Smart Speakern, dass Wake-Word-Systeme unbeabsichtigt ausgelöst werden können, etwa durch ähnlich klingende Wörter oder akustische Fehlinterpretationen (Schönherr et al., 2022; Chen et al., 2021). Das Wake-Word-Gating reduziert daher zwar die Menge potenziell sensibler Rohdaten und begrenzt den Energiebedarf, muss aber durch transparente Statusanzeigen, Protokollierung und robuste Fehlerbehandlung ergänzt werden. Für die Spracherkennung kommen lokal ausführbare Whisper-Modelle beziehungsweise Whisper-kompatible Laufzeiten in Betracht. Whisper wurde auf 680.000 Stunden multilingualer und multitaskfähiger Sprachdaten trainiert und ist auf robuste Spracherkennung unter variierenden Bedingungen ausgelegt (Radford et al., 2023). Lokale Implementierungen wie whisper.cpp unterstützen darüber hinaus quantisierte Modelle, wodurch Speicherbedarf und Laufzeitverhalten für Edge-Systeme optimiert werden können.

Die lokale LLM-Inferenz mit Gemma 4 ergänzt die STT-Pipeline um semantische Verarbeitung. Nach der Transkription können Nutzerintention, Kontext, Rollenprofil und gewünschte Aktion lokal bestimmt werden. Die bereits vorgesehene Profil- und Rollenarchitektur ist dabei ein wesentliches Element: Ein Profil kann mehrere Rollen enthalten, etwa Professor, Geschäftsführer oder private Rolle. Jede Rolle kann eigene Schreibstile, Datenhaltungsregeln, Wissensbasen, E-Mail-Absender, Freigabestufen und Cloud-Zugriffsrechte besitzen. Damit wird verhindert, dass private, wissenschaftliche und geschäftliche Kontexte unkontrolliert vermischt werden. Theoretisch lässt sich diese Architektur mit Arbeiten zu kontextabhängigen Datenschutz-Normen und Informationsflüssen in Smart Assistants verbinden. Forschung zu Smart Home Personal Assistants zeigt, dass Datenschutzentschei-

dungen nicht losgelöst vom Nutzungskontext betrachtet werden sollten, sondern davon abhängen, welche Information von wem, über wen, an wen und unter welchen Bedingungen weitergegeben wird (Abdi et al., 2021). Entsprechend kann die Rollenarchitektur als technische Operationalisierung einer Policy-Schicht verstanden werden, die Informationsflüsse kontextabhängig erlaubt, einschränkt oder verhindert. Arbeiten zur Governance von Information Sharing in Smart Assistants zeigen zudem, dass explizite Privacy-Norm-Repräsentationen und kontextabhängige Entscheidungsmodelle geeignet sind, um das Verhalten persönlicher Assistenten stärker an Datenschutzpräferenzen auszurichten (Zhan et al., 2022). Nicht nur die Modellantwort ist daher relevant, sondern auch die Frage, welche Daten überhaupt verarbeitet, gespeichert, synchronisiert oder an externe Systeme übergeben werden dürfen.

3.3 Privacy-by-Design

Die Datenschutzarchitektur folgt einem Privacy-by-Design-Ansatz und verbindet lokale Verarbeitung mit für Menschen unmittelbar wahrnehmbaren Transparenzsignalen. Da sprachbasierte Assistenzsysteme im Alltag häufig in geteilten sozialen Räumen verwendet werden, reicht eine rein technische Absicherung nicht aus; vielmehr müssen Nutzende und Umstehende erkennen können, wann ein Gerät lediglich betriebsbereit ist, wann es auf ein Wake Word wartet, wann es tatsächlich aufnimmt und wann eine lokale Verarbeitung erfolgt. Empirische Untersuchungen zu Smart-Speaker-Nutzenden verdeutlichen, dass Datenschutzbedenken, Kontrollmöglichkeiten und Transparenz über die Geräteaktivität wesentlich für die Wahrnehmung solcher Systeme sind (Malkin et al., 2019). Zusätzlich ist zu berücksichtigen, dass Wake-Word-Systeme unbeabsichtigt ausgelöst werden können, etwa durch ähnlich klingende Sprachmuster oder akustische Fehlinterpretationen (Schönherr et al., 2022). Vor diesem Hintergrund ist für das NeuroMate-System eine direkt am Gerät integrierte LED-Statusanzeige vorgesehen, beispielsweise als LED-Statusleiste oder als einzelne mehrfarbige Status-LED. Diese Anzeige soll nicht lediglich eine nachträgliche softwareseitige Visualisierung sein, sondern möglichst eng mit dem tatsächlichen Zustand der Audioaufnahme- und Verarbeitungspipeline gekoppelt werden. Dadurch wird unmittelbar sichtbar, ob das Gerät passiv auf ein Wake Word wartet, aktiv aufnimmt, lokal transkribiert, inferiert oder ein Protokollereignis schreibt. Ergänzend kann ein akustischer Indikator vorgesehen werden, etwa ein kurzer Signalton bei Aktivierung oder Deaktivierung der Aufnahme.

Zusätzlich wird ein lokales Protokoll geführt, das relevante Systemzustände und Verarbeitungsschritte nachvollziehbar dokumentiert, ohne unnötig sensible Inhalte zu speichern. Diese lokale Protokollierung entspricht der Grundidee des Edge Computing, Daten und Verarbeitung möglichst nahe am Entstehungsort zu halten, um Latenz, Netzwerkabhängigkeit und externe Datenweitergabe zu reduzieren (Shi et al., 2016; Satyanarayanan, 2017;

Yu et al., 2018). Das Protokoll kann optional über eine Companion-App auf dem Smartphone gespiegelt werden. Die App dient damit als erweiterte Status-, Kontroll- und Demonstrationsoberfläche, ist jedoch keine zwingende Voraussetzung für den Standalone-Betrieb des Geräts. Für Vortrags- und Demonstrationssituationen eröffnet diese Architektur mehrere Möglichkeiten: Der Status der Companion-App kann beispielsweise auf einem Mac-Bildschirm gespiegelt werden, während das physische Gerät parallel über eine Kamera gezeigt wird; alternativ kann ein vorbereitetes Demonstrationsvideo die Interaktion zwischen Gerät, LED-Statusanzeige und App-Oberfläche veranschaulichen. Neben der Smartphone-App wäre auch eine direkte Statusausgabe über den vorhandenen HDMI-Anschluss der Hardware möglich. Ein eInk-Display wird demgegenüber nicht als notwendiger Bestandteil der aktuellen Architektur verstanden, sondern als optionales Roadmap-Feature für spätere Ausbaustufen, etwa für besonders stromsparende, persistente Statusanzeigen direkt am Gerät. Die Statusinformationen sind damit nicht nur Bedienkomfort, sondern ein zentrales Element sozialer Transparenz, technischer Nachvollziehbarkeit und datenschutzorientierter Systemgestaltung.

Für Umstehende ist zusätzlich ein Ausblick auf **Tangible Privacy** sinnvoll. Forschung zu Smart Voice Assistants zeigt, dass physische und greifbare Datenschutzkontrollen von Umstehenden als vertrauenswürdiger und verständlicher wahrgenommen werden können als rein softwarebasierte Mechanismen (Ahmad et al., 2022). Auch kontextadaptive Privacy Controls, etwa durch Blickrichtung, Lautstärke oder andere interpersonale Kommunikationssignale, wurden als Möglichkeit untersucht, Mikrofonaktivierung selektiver zu gestalten (Mhaidli et al., 2020). Für NeuroMate könnten daraus konkrete Opt-out-Mechanismen entstehen: ein physischer Privacy-Schalter, ein sichtbarer „Private Mode“, lokale Sperrregeln für Mehrpersonen-Transkripte, ein Raum-Beacon, ein Tischaufsteller mit Status-LED oder App-basierte Sitzungsinformationen für Meetings.

Auditierbarkeit und Security werden durch lokale Protokolle, Versionierung und reproduzierbare Deployments gestützt. Docker dient dabei nicht nur der Bequemlichkeit, sondern der wissenschaftlichen und technischen Nachvollziehbarkeit. Reproduzierbarkeit entsteht jedoch nicht allein durch Containerisierung, sondern durch konsequent versionierte Images, gepinnte Abhängigkeiten, dokumentierte Build-Parameter, Hashes, Modellversionen und nachvollziehbare Konfigurationsdateien. Nüst et al. (2020) zeigen, dass gut dokumentierte Dockerfiles und reproduzierbare Container für wissenschaftliche Workflows einen erheblichen Beitrag zur Transparenz und Nachprüfbarkeit leisten können.

Für Updates sind signierte, versionierte Firmware- und Softwarepakete mit Rollback-Fähigkeit vorgesehen. Gerade IoT- und Edge-Systeme benötigen sichere Update-Prozesse, weil sie langfristig betrieben werden, im Feld schwer zugänglich sind und zugleich regel-

mäßig Sicherheitskorrekturen benötigen. El Jaouhari und Bouvet (2022) beschreiben Over-the-Air-Firmware-Updates als zentrales Sicherheitsinstrument für IoT-Geräte und diskutieren Herausforderungen wie Integrität, Authentizität, Angriffsszenarien und Update-Mechanismen. Für NeuroMate bedeutet dies, dass Firmware-Updates nicht nur funktional, sondern kryptografisch abgesichert, versioniert, überprüfbar und bei Fehlern zurücksetzbar sein sollten.

Insgesamt entsteht eine Architektur, in der lokale KI-Inferenz, spezialisierte Edge-Beschleunigung und datenschutzorientiertes Systemdesign zusammengeführt werden. Das RTX-5090-Host-System fungiert als leistungsfähige lokale Entwicklungs- und Evaluationsplattform. Der Banana Pi M7 mit RK3588, NVMe-Speicher und Hailo-10H-Modul bildet die mobile Edge-Zielplattform. Whisper-basierte lokale Spracherkennung entfernt Cloud-Abhängigkeiten im Audioeingang. Gemma 4 über Docker und Ollama stellt die lokale semantische Inferenz bereit, wobei größere Modelle auf dem Host und kleinere beziehungsweise quantisierte Varianten auf dem Edge-Gerät betrieben werden können. Hardware-Indikatoren, Wake-Word-Gating, lokale Verarbeitung, Tangible-Privacy-Mechanismen, versionierte Container, lokale Audit-Logs und signierte Updates bilden gemeinsam eine technische Grundlage für ein tragbares KI-System, das nicht nur leistungsfähig, sondern auch nachvollziehbar, kontrollierbar und sozial transparent gestaltet ist.

4. Evaluation und Trade-Offs

Die Evaluation des NeuroMate-Prototyps erfolgt in einem mehrstufigen Testaufbau, der technische Leistungsfähigkeit, datenschutzbezogene Robustheit und alltagspraktische Nutzbarkeit gemeinsam betrachtet. Ziel ist nicht nur die Messung isolierter Modellgüte, sondern die Bewertung der gesamten Edge-Pipeline vom Wake-Word-Gating über Audioaufnahme, Speech-to-Text-Verarbeitung und lokale LLM-Inferenz bis hin zur Statusanzeige und Protokollierung. Als zentrale Messgrößen werden Latenz, Echtzeitfähigkeit, Erkennungsqualität, Energieverbrauch, thermisches Verhalten, Fehlaktivierungen und Robustheit in unterschiedlichen akustischen Umgebungen erfasst. Für die Spracherkennung werden klassische Metriken wie Word Error Rate, Match Error Rate und Word Information Lost herangezogen, da sie unterschiedliche Perspektiven auf die Transkriptionsqualität eröffnen (Morris et al., 2004). Ergänzend sollten anwendungsnahe Metriken aufgenommen werden, etwa Intent-Erkennungsrate, Fehler bei der Sprecherzuordnung, Anteil nicht verwertbarer Transkriptsegmente und Zeit bis zur nutzbaren Zusammenfassung.

Der Testaufbau umfasst mehrere realitätsnahe Szenarien: ruhiges Büro, Seminarraum, Vorlesungssituation, Besprechungsraum, Flur- oder Cafeteriaumgebung sowie mobile

Nutzung. In jedem Szenario werden definierte Distanzen zwischen Sprecher, Gerät und Nebensprechern geprüft. Besondere Aufmerksamkeit verdient die Fernfeldproblematik: In einer Vorlesung kann der primäre Sprecher räumlich weiter vom Gerät entfernt stehen, während Studierende näher am Mikrofon sitzen. In diesem Fall besteht das Risiko, dass Kommentare oder Nebengespräche der Umstehenden akustisch dominanter erfasst werden als der eigentlich intendierte Sprechtext. Aus technischer Sicht handelt es sich dabei nicht nur um ein Qualitätsproblem, sondern auch um ein Datenschutzproblem, weil das System möglicherweise gerade diejenigen Personen besser transkribiert, die nicht primäre Nutzende des Systems sind. Forschung zu distant multi-microphone conversational ASR, etwa im Rahmen der CHiME-5-Challenge, zeigt, dass natürliche Mehrpersonengespräche in realen Umgebungen mit Distanz, Hall, Überlappungen und Mikrofonpositionen erhebliche Herausforderungen für Spracherkennung und Signalverarbeitung darstellen (Barker et al., 2018).

Eine weitere Evaluationsdimension betrifft parallele Sprecher und Sprecherzuordnung. Speaker Diarization adressiert die Frage, „wer wann gesprochen hat“, und ist insbesondere bei Mehrpersonenszenarien, Meetings und Vorlesungsinteraktionen relevant (Anguera et al., 2012; Park et al., 2022). Für NeuroMate sollten daher nicht nur Transkriptionsfehler gemessen werden, sondern auch Diarization Error Rate, Overlap-Erkennung und Fehler bei der Zuordnung zwischen aktivem Sprecher und Umgebungsstimmen. Damit wird überprüfbar, ob das System den intendierten Nutzenden priorisiert oder ob es unbeteiligte Dritte überproportional erfasst.

Ein pragmatischer Lösungsansatz besteht darin, den primären Sprecher über ein Headset oder Ansteckmikrofon als bevorzugte Audioquelle zu koppeln. Ein per Bluetooth angebundenes Headset könnte insbesondere in Vorlesungs- und Vortragssituationen als primärer Kanal dienen, während das Gerät selbst nur ergänzende Umgebungsinformationen verarbeitet oder bewusst auf diese verzichtet. Technisch führt dies jedoch in Richtung Multi-Mikrofon-Processing, Quellenpriorisierung, Beamforming, Sprecherlokalisierung und gegebenenfalls Sprechertrennung. Diese Verfahren sollten nicht als triviale Zusatzfunktion behandelt werden, sondern als eigenständige Ausbaustufe. Für die erste Evaluation ist daher zu unterscheiden zwischen dem Standalone-Betrieb mit geräteeigenem Mikrofon, dem Betrieb mit bevorzugtem externem Mikrofon und einer späteren Multi-Source-Architektur.

Der zentrale Trade-off liegt zwischen maximalem On-Device-Datenschutz und praktischen Ressourcenbeschränkungen. Lokale Inferenz reduziert Cloud-Abhängigkeiten und externe Datenflüsse, erhöht aber Anforderungen an sowohl die Rechenleistung, die Energieversorgung, die Kühlung, die Modellkompression als auch an die Systemintegration. Edge-AI-

Forschung beschreibt diesen Zielkonflikt zwischen Latenz, Bandbreite, Datenschutz, Modellgröße und verfügbarer Hardware als Kernproblem lokaler Inferenzarchitekturen (Chen & Ran, 2019; Zhou et al., 2019). Für NeuroMate bedeutet dies: Die beste datenschutztechnische Lösung ist nicht automatisch die leistungsfähigste Lösung, und die leistungsfähigste lokale Modellvariante ist nicht automatisch die mobilste. Die Evaluation muss daher transparent ausweisen, welche Modellgröße, Quantisierung, Energieaufnahme und Antwortqualität in welchem Nutzungsszenario erreicht wird.

5. Pragmatische Mindeststandards für Wearable-KI

Aus dem Prototyp lassen sich pragmatische Mindeststandards für tragbare KI-Systeme ableiten. Diese Standards sollten nicht nur Modellleistung bewerten, sondern die gesamte soziotechnische Architektur: lokale Verarbeitung, Transparenz, Bystander Privacy, Auditierbarkeit, Updatefähigkeit und Datenhaltungsgrenzen. Daraus ergibt sich der Vorschlag eines Trustworthy Wearable AI Labels, das Edge-Privacy und Transparenz sichtbar zertifiziert. Die Idee eines solchen Labels ist wissenschaftlich anschlussfähig, da Zertifizierungslabels für Trustworthy AI empirisch als Mittel untersucht wurden, um Vertrauenswürdigkeit gegenüber Endnutzenden verständlicher zu kommunizieren (Scharowski et al., 2023).

Ein solches Label könnte in Mindestkriterien gegliedert werden: Erstens muss das System offenlegen, ob Audio-, Text- oder Sensordaten lokal verarbeitet oder an externe Dienste übertragen werden. Zweitens braucht es eine unmittelbar sichtbare Statusanzeige, etwa über eine hardware-nahe LED-Statusleiste, die Aufnahme, Transkription, Inferenz und Protokollierung unterscheidbar macht. Drittens müssen Wake-Word-Gating, Fehlaktivierungen und Opt-out-Mechanismen für Umstehende dokumentiert werden. Viertens sind klare Datenhaltungsgrenzen erforderlich: Roh-Audio sollte standardmäßig nicht dauerhaft gespeichert werden; Transkripte und Zusammenfassungen benötigen definierte Speicherfristen, Rollenbezug und Löschmechanismen. Fünftens sollten Modelle, Datengrundlagen, Evaluationsszenarien und Grenzen des Systems dokumentiert werden. Hierfür bieten Model Cards, Datasheets for Datasets und interne Auditberichte etablierte Vorbilder für transparente KI-Dokumentation (Mitchell et al., 2019; Gebru et al., 2021; Raji et al., 2020).

Aus gutachterlicher Perspektive ließe sich dieses Label an drei Prüfdimensionen anschließen: Informationstechnologie, Informationsmanagement und KI-Transformation sowie Prozessdatenverarbeitung, Echtzeitsysteme und dezentrale Edge-Inferenz. Damit würde Wearable-KI nicht nur als Produktkategorie, sondern als prüfbares System aus Hardware,

Software, Datenflüssen, Nutzerinteraktion und Governance verstanden. Trustworthy Wearable AI wäre dann kein Marketingversprechen, sondern eine nachvollziehbare Konformitäts- und Qualitätsaussage.

6. Fazit und Ausblick

Der NeuroMate-Prototyp zeigt, dass Edge-AI für tragbare Assistenzsysteme nicht nur eine technische Option, sondern eine datenschutzstrategische Alternative zu cloudzentrierten KI-Wearables darstellt. Die Verfügbarkeit leistungsfähiger Platinen, lokaler KI-Beschleuniger, quantisierter Sprachmodelle und containerisierter Deployment-Umgebungen verschiebt die Grenze dessen, was ohne permanente Cloud-Anbindung möglich ist. Die eigentliche Herausforderung liegt zunehmend weniger in der reinen Rechenleistung, sondern in der zuverlässigen Ansteuerung der Komponenten, der Modellintegration, der Energieeffizienz, der transparenten Nutzerführung und der Governance des Gesamtsystems.

Für Bildung, Beschaffung und Standardisierung ist diese Entwicklung besonders relevant. Bildungseinrichtungen, Unternehmen und öffentliche Organisationen werden künftig nicht nur fragen müssen, welches KI-System die beste Antwort liefert, sondern wo Daten verarbeitet werden, welche Personen miterfasst werden, welche Protokolle entstehen und wie die Einhaltung von Datenschutz- und Transparenzanforderungen nachgewiesen wird. Forschung zu Trustworthy AI betont, dass Vertrauenswürdigkeit nicht allein aus Modellgüte entsteht, sondern aus einem Zusammenspiel von Transparenz, Verantwortung, Datenschutz, Robustheit und Nachvollziehbarkeit (Jobin et al., 2019; Thiebes et al., 2021). Organisatorische AI-Governance muss daher Daten, Modelle und KI-Systeme gemeinsam betrachten und Verantwortlichkeiten entlang des gesamten Lebenszyklus definieren (Schneider et al., 2023).

Der Ausblick ist entsprechend klar: Edge-AI wird leistungsfähiger, lokaler und stärker regulierungsrelevant. Für tragbare KI-befähigte Geräte, also so. “Wearable-KI” entsteht daraus ein Pionierfeld der Forschung, in welchem sowohl die technische Innovation, als auch europäische Datenschutzprinzipien und sachverständige Prüfbarkeit zusammengeführt werden können. Vollständige DSGVO-Kompatibilität bleibt anspruchsvoll, erscheint aber gerade durch Edge-Inferenz, lokale Protokollierung, transparente Statusanzeigen und klare Datenhaltungsgrenzen realistischer als bei cloudzentrierten Always-on-Geräten. Die weitere Entwicklung sollte daher nicht nur auf größere Modelle zielen, sondern auf überprüfbare, reproduzierbare und sozial transparente Systeme.

Literature

Abdi, N., Zhan, X., Ramokapane, K. M., & Such, J. M. (2021). Privacy norms for smart home personal assistants. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (Article 558, pp. 1–14). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445122

Ahmad, I., Akter, T., Buher, Z., Farzan, R., Kapadia, A., & Lee, A. J. (2022). Tangible privacy for smart voice assistants: Bystanders' perceptions of physical device controls. Proceedings of the ACM on Human-Computer Interaction, 6(CSCW2), Article 364. https://doi.org/10.1145/3555089

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., & Vinyals, O. (2012). Speaker diarization: A review of recent research. IEEE Transactions on Audio, Speech, and Language Processing, 20(2), 356–370. https://doi.org/10.1109/TASL.2011.2125954

Barker, J., Watanabe, S., Vincent, E., & Trmal, J. (2018). The fifth "CHiME" speech separation and recognition challenge: Dataset, task and baselines. Proceedings of Interspeech 2018, 1561–1565. https://doi.org/10.21437/Interspeech.2018-1768

Benazir, A., & Lin, F. X. (2025). Safeguarding privacy in edge speech understanding with tiny foundation models. Proceedings of the Tenth ACM/IEEE Symposium on Edge Computing. https://doi.org/10.1145/3769102.3770609

Cavoukian, A. (2009). Privacy by design: The 7 foundational principles. Information and Privacy Commissioner of Ontario.

Chen, J., & Ran, X. (2019). Deep learning with edge computing: A review. Proceedings of the IEEE, 107(8), 1655–1674. https://doi.org/10.1109/JPROC.2019.2921977

Chen, Y., Bai, Y., Mitev, R., Wang, K., Sadeghi, A.-R., & Xu, W. (2021). FakeWake: Understanding and mitigating fake wake-up words of voice assistants. In Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (pp. 1861–1883). Association for Computing Machinery. https://doi.org/10.1145/3460120.3485365

Cheng, P., & Roedig, U. (2022). Personal voice assistant security and privacy—A survey. Proceedings of the IEEE, 110(4), 476–507. https://doi.org/10.1109/JPROC.2022.3153167

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. Advances in Neural Information Processing Systems, 36, 10088–10115. https://doi.org/10.52202/075280-0441

Dey, A. K., Abowd, G. D., & Salber, D. (2001). A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. Human–Computer Interaction, 16(2–4), 97–166. https://doi.org/10.1207/S15327051HCI16234_02

Dey, S. (2024). A task-agnostic distillation method for generating compact yet generalized models: Preliminary proof-of-concept results demonstrating the potential of edge foundation models. Proceedings of the Workshop on Edge and Mobile Foundation Models.

Doherty, C., Baldwin, M., Lambe, R., Altini, M., & Caulfield, B. (2025). Privacy in consumer wearable technologies: A living systematic analysis of data policies across leading manufacturers. npj Digital Medicine, 8(1). https://doi.org/10.1038/s41746-025-01757-1

El Jaouhari, S., & Bouvet, E. (2022). Secure firmware Over-The-Air updates for IoT: Survey, challenges, and discussions. Internet of Things, 18, Article 100508. https://doi.org/10.1016/j.iot.2022.100508

European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016. Official Journal of the European Union, L 119, 1–88.

Europäische Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

Fang, C. M., Samaradivakara, Y., Maes, P., & Nanayakkara, S. (2025). Mirai: A wearable proactive AI "inner-voice" for contextual nudging. In Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (Article 399, pp. 1–9). Association for Computing Machinery. https://doi.org/10.1145/3706599.3719881

Frantar, E., Ashkboos, S., Hoefler, T., & Alistarh, D. (2023). GPTQ: Accurate post-training quantization for generative pre-trained transformers. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2210.17323

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., III, & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92. https://doi.org/10.1145/3458723

Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M. W., & Keutzer, K. (2022). A survey of quantization methods for efficient neural network inference. In Low-Power Computer Vision (pp. 291–326). CRC Press. https://doi.org/10.1201/9781003162810-13

Gondi, S. (2021). Energy-accuracy trade-off of ASR with a typical transformer-based speech recognition model on an edge device. Sustainability.

Hernández Acosta, L., & Reinhardt, D. (2022). A survey on privacy issues and solutions for voice-controlled digital assistants. Pervasive and Mobile Computing, 80, Article 101523. https://doi.org/10.1016/j.pmcj.2021.101523

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1, 389–399. https://doi.org/10.1038/s42256-019-0088-2

Koelle, M., Ananthaswamy, W., & Boll, S. (2020). Tangible privacy: Towards user-centric sensor designs for bystander privacy. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW2), 1–28. https://doi.org/10.1145/3415187

Kröger, J. L., Gellrich, L., Pape, S., Brause, S. R., & Ullrich, S. (2022). Personal information inference from voice recordings: User awareness and privacy concerns. Proceedings on Privacy Enhancing Technologies, 2022(1), 6–27. https://doi.org/10.2478/popets-2022-0002

Li, J., Chen, C., Rahimi Azghadi, M., Ghodosi, H., Pan, L., & Zhang, J. (2023). Security and privacy problems in voice assistant applications: A survey. Computers & Security, 134, Article 103448. https://doi.org/10.1016/j.cose.2023.103448

Malkin, N., Deatrick, J., Tong, A., Wijesekera, P., Egelman, S., & Wagner, D. (2019). Privacy attitudes of smart speaker users. Proceedings on Privacy Enhancing Technologies, 2019(4), 250–271. https://doi.org/10.2478/popets-2019-0068

Mhaidli, A. H., Venkatesh, M. K., Zou, Y., & Schaub, F. (2020). Listen only when spoken to: Interpersonal communication cues as smart speaker privacy controls. Proceedings on Privacy Enhancing Technologies, 2020(2), 251–270. https://doi.org/10.2478/popets-2020-0026

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220–229). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287596

Morris, A. C., Maier, V., & Green, P. (2004). From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition. Proceedings of Interspeech 2004, 2765–2768. https://doi.org/10.21437/Interspeech.2004-668

Musch, S., Borrelli, M. C., & Kerrigan, C. (2024). Bridging compliance and innovation: A comparative analysis of the EU AI Act and GDPR for enhanced organisational strategy. Journal of Data Protection & Privacy, 7(1), 14–40. https://doi.org/10.69554/FWHU3837

Nüst, D., Sochat, V., Marwick, B., Eglen, S. J., Head, T., Hirst, T., & Evans, B. D. (2020). Ten simple rules for writing Dockerfiles for reproducible data science. PLOS Computational Biology, 16(11), Article e1008316. https://doi.org/10.1371/journal.pcbi.1008316

Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., & Narayanan, S. (2022). A review of speaker diarization: Recent advances with deep learning. Computer Speech & Language, 72, Article 101317. https://doi.org/10.1016/j.csl.2021.101317

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. Proceedings of the 40th International Conference on Machine Learning (ICML). https://doi.org/10.48550/arXiv.2212.04356

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 33–44). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372873

Satyanarayanan, M. (2017). The emergence of edge computing. Computer, 50(1), 30–39. https://doi.org/10.1109/MC.2017.9

Scharowski, N., Benk, M., Kühne, S. J., Wettstein, L., & Brühlmann, F. (2023). Certification labels for trustworthy AI: Insights from an empirical mixed-method study. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 248–260). Association for Computing Machinery. https://doi.org/10.1145/3593013.3593994

Schneider, J., Abraham, R., Meske, C., & vom Brocke, J. (2023). Artificial intelligence governance for businesses. Information Systems Management, 40(3), 229–249. https://doi.org/10.1080/10580530.2022.2085825

Schönherr, L., Zeiler, M., Holz, T., & Kolossa, D. (2022). Exploring accidental triggers of smart speakers. Computer Speech & Language, 73, 101328. https://doi.org/10.1016/j.csl.2021.101328

Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. IEEE Internet of Things Journal, 3(5), 637–646. https://doi.org/10.1109/JIOT.2016.2579198

Somvanshi, S. (2025). From TinyML to TinyDL: A comprehensive survey of architectural innovations, hardware platforms, model optimization techniques, and emerging directions including neuromorphic computing and edge-native foundation models. ACM Computing Surveys.

Takhshid, Z. (2024). Wearable AI, bystander notice, and the question of privacy frictions. Boston University Law Review, 104(4), 1087-1106.

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. Electronic Markets, 31, 447–464. https://doi.org/10.1007/s12525-020-00441-4

Torkamani, M. J. (2025). ASTA: An adaptive speech-to-action solution that dynamically routes voice commands between edge and cloud inference to balance performance and system resource utilization. ArXiv.

Yu, W., Liang, F., He, X., Hatcher, W. G., Lu, C., Lin, J., & Yang, X. (2018). A survey on the edge computing for the Internet of Things. IEEE Access, 6, 6900–6919. https://doi.org/10.1109/ACCESS.2017.2778504

Zhan, X., Sarkadi, S., Criado, N., & Such, J. M. (2022). A model for governing information sharing in smart assistants. In AIES 2022: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (pp. 845–855). Association for Computing Machinery. https://doi.org/10.1145/3514094.3534129

Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., & Zhang, J. (2019). Edge intelligence: Paving the last mile of artificial intelligence with edge computing. Proceedings of the IEEE, 107(8), 1738–1762. https://doi.org/10.1109/JPROC.2019.2918951

Internet sources

NVIDIA. (n.d.). GeForce RTX 5090-Grafikkarten. Retrieved April 29, 2026, from https://www.nvidia.com/de-de/geforce/graphics-cards/50-series/rtx-5090/

Banana Pi Wiki. (2024, April 13). Banana Pi BPI-M7. Retrieved April 29, 2026, from https://wiki.banana-pi.org/Banana_Pi_BPI-M7

Hailo. (n.d.). Hailo-10H M.2 AI acceleration module. Retrieved April 29, 2026, from https://hailo.ai/products/ai-accelerators/hailo-10h-m-2-ai-acceleration-module/

Google AI for Developers. (2026, April 16). Gemma 4-Modelle – Übersicht. https://ai.google.dev/gemma/docs/core?hl=de

Limitless AI. (2025, December 5). Privacy policy. https://www.limitless.ai/privacy-policy

Chokkattu, J. (2024, April 11). Humane Ai Pin review: Too clunky, too limited. WIRED. https://www.wired.com/review/humane-ai-pin/