

Transparente Mensch-Fahrzeug-Kommunikation

Wie erklärbare Kommunikation Vertrauen gewinnen und regulatorische Freigabe für höhere Autonomiestufen erreichen kann

fir an der RWTH Aachen

DiSer Hub

Inhalt

Einleitung 3

Von der KI-Entscheidung zur vertrauensbildenden Fahrzeugkommunikation 5

- Technische Grundlagen und Methoden der Explainable AI im Fahrzeug 5
- Zukunftsfähige Human-Machine-Interfaces 6
- Implementierungsbeispiele entlang der SAE-Level 7

Regulatorischer Rahmen und juristische Implikationen 11

Fazit und Ausblick 13

Literaturverzeichnis 14

Abkürzungsverzeichnis 15

Impressum

Autoren
Jonas Schatton
FIR e. V. an der RWTH Aachen

Gustav Rake
FIR e. V. an der RWTH Aachen

Satz und Design
FIR e. V. an der RWTH Aachen

Bildnachweis
S. 1: © TSViPhoto – stock.adobe.com

KI-Transparenzhinweis:
Bei der Erstellung dieser Publikation wurden OpenAI GPT-6 und GPT-5.6 Sol zur Unterstützung bei der Literaturrecherche, Recherche von Praxisbeispielen sowie der Texterstellung und sprachlichen Überarbeitung verwendet. Der Text wurde anschließend noch von den Autoren geprüft und überarbeitet.

Zitierempfehlung:
Schatton, J.; Rake, G.: [Whitepaper] Transparente Mensch-Fahrzeug-Kommunikation: Wie erklärbare Kommunikation Vertrauen gewinnen und regulatorische Freigabe für höhere Autonomiestufen erreichen kann. FIR e. V. an der RWTH Aachen, Aachen 2026. DOI: 10.60571/4467

Lizenzhinweis

Open Access: Dieses Whitepaper wird unter der Creative-Commons-Lizenz „Share alike – Weitergabe unter gleichen Bedingungen 4.0 International" (CC BY-SA 4.0) veröffentlicht.

DOI: 10.60571/4467

Projekt DiSerHub
FIR e. V. an der RWTH Aachen
Campus-Boulevard 55 I 52074 Aachen

E-Mail: projekt-DiSerHub@fir.rwth-aachen.de
diserhub.de

Stand: 11.2026

Einleitung: Wenn Fahrzeuge entscheiden, aber niemand versteht, wie und warum

San Francisco gilt seit langem als Mekka des autonomen Fahrens. Unternehmen wie *Waymo* betreiben dort bereits seit mehreren Jahren zahlreiche Robotaxis und nutzen die Stadt als urbanes Reallabor. Schon heute zeigt sich, dass autonome Fahrzeuge im Vergleich zu menschlichen Fahrer*innen eine messbar geringere Unfallrate aufweisen[1].

Dennoch stockt das Vertrauen der Öffentlichkeit – die statistisch starke Sicherheitsbilanz allein scheint nicht zu genügen. Besonders kritisch werden Situationen wahrgenommen, in denen autonome Fahrzeuge aufgrund von Systemüberlastungen abrupt stoppen, auf Kreuzungen stehen bleiben oder Rettungswege blockieren. Solche Vorfälle erschüttern das Vertrauen der Fahrgäste und verstärken den Eindruck, das Verhalten der Systeme nicht zuverlässig vorhersehen zu können[2]. Während sich die Entscheidungen menschlicher Fahrer*innen zumindest intuitiv deuten lassen, bleiben die Entscheidungslogiken autonomer Systeme den Nutzer*innen bislang weitgehend verborgen.

Mit der Einführung autonomer Fahrzeuge vollzieht sich ein grundlegender Wandel der Rolle der Fahrzeugnutzer*innen: Der Mensch wandelt sich von der aktiven Fahrzeugführung zur passiven Supervisor-Instanz eines komplexen technischen Systems, dessen Verhalten er*sie beobachten, bewerten und im Zweifel einordnen muss. Die Akzeptanz autonomer Fahrzeuge hängt daher nicht allein von ihrer objektiven Funktionssicherheit ab, sondern in hohem Maße auch davon, ob ihr Verhalten für die Nutzer*innen verständlich und berechenbar erscheint. Mit dem Einzug höherer Autonomiestufen in den stark regulierten deutschen und europäischen Straßenverkehr gewinnt deshalb eine zentrale Frage an Dringlichkeit:

Wie können Fahrzeuge zukünftig kommunizieren, um Vertrauen in ihre Entscheidungen und Reaktionen zu schaffen?

Die Nachvollziehbarkeit einer Systementscheidung hängt maßgeblich von ihrer Transparenz ab. Moderne autonome Fahrzeuge erfassen ihre Umgebung über ein dichtes Netz aus Sensoren, Kameras und Radarsystemen und planen ihre Reaktionen auf der Basis komplexer KI-Modelle. Gerade die Architektur dieser Systeme erschwert jedoch ihre Transparenz. Die internen Entscheidungsprozesse basieren auf mathematischen Berechnungen und sind nicht unmittelbar in Worten und Sätzen nachvollziehbar. Für den Menschen agieren sie daher weitgehend als „Black Box“[3]. Die Ursache hinter einer Systementscheidung, wie einem abrupten Stopp im fließenden Verkehr, bleibt dadurch unklar. Mit Explainable AI (XAI) kann diese Intransparenz überwunden werden: Sie dient dazu, die verborgene Logik algorithmischer Entscheidungen in für Menschen verständliche Informationen zu übersetzen und damit die Lücke zwischen maschineller Verarbeitung und menschlichem Verständnis zu schließen.

Hierdurch ändert sich auch die Gestaltung der Human-Machine-Interfaces (HMI). Während konventionelle Konzepte vor allem den aktuellen Systemstatus anzeigen, rücken bei zukunftsfähigen HMIs die Erklärung und Kontextualisierung des Fahrzeugver-

[1] s. Kusano et al. 2025, S. 20f.
[2] s. Deng u. He 2026, S. 28
[3] s. Ali u. Latif 2026, S. 1

haltens in den Mittelpunkt. Sie informieren nicht nur über das „Was“, etwa eine eingeleitete Bremsung, sondern vermitteln zugleich das „Warum“, beispielsweise die Erkennung von Fußgänger*innen, die die Fahrbahn überqueren. Die Übersetzung der internen Systemwahrnehmung in Schrift und Sprache kann dabei über verschiedene Formen und Sinne erfolgen. Visuelle Hervorhebungen relevanter Objekte, adaptive Lichtsignale, akustische Hinweise oder haptisches Feedback machen sichtbar und spürbar, welche Faktoren das Fahrzeug erkannt und bei seiner Entscheidung berücksichtigt hat. Damit werden die internen Bewertungen des Algorithmus unmittelbar erkennbar und für die Nutzer*innen nachvollziehbar.[4]

Nicht nur für die individuelle Nutzer*innenakzeptanz, auch für die Zulassung automatisierter Systeme ist die Transparenz von KI-Prozessen von fundamentaler Bedeutung. Da die Verantwortung zunehmend vom Menschen auf das Fahrzeug und die Software übergeht, also statt von der menschlichen Handlungskontrolle von der algorithmischen Determinierung abhängt, wird die Nachvollziehbarkeit der Algorithmen auch für die regulatorischen Anforderungen immer zentraler. Für Systeme höherer Autonomiestufen (SAE-Level 3 bis 5), in denen das Fahrzeug die Fahraufgabe zeitweise oder dauerhaft ohne menschliche Überwachung übernimmt, verschärfen sich die Zulassungshürden weiter. Auch die Rechtsprechung steht damit vor neuen Herausforderungen in der Haftungs- und Forensikfrage.[5]

Die nachfolgenden Kapitel zeigen, wie sich die Entscheidungsprozesse autonomer Systeme technisch erklären, über geeignete Human-Machine-Interfaces verständlich kommunizieren und entlang der unterschiedlichen SAE-Level praktisch umsetzen lassen. Abschließend wird deutlich, welche regulatorischen und haftungsrechtlichen Anforderungen sich daraus für die Entwicklung und Zulassung autonomer Fahrzeuge ergeben.

[4] s. Luo et al. 2025, S. 3

[5] s. Tanenblatt et al. 2025, S. 26f.

Von der KI-Entscheidung zur vertrauensbildenden Fahrzeugkommunikation

Technische Grundlagen und Methoden der Explainable AI im Fahrzeug

Die technische Erklärbarkeit autonomer Fahrentscheidungen hängt maßgeblich von der zugrunde liegenden KI-Architektur ab. Moderne Wahrnehmungs- und Planungssysteme funktionieren häufig nach einem End-to-End-Ansatz. Dabei werden Sensordaten aus Kamera-, Radar- und weiteren Sensorsystemen unmittelbar verarbeitet und direkt in Fahrentscheidungen beziehungsweise Steuerungsbefehle übersetzt[6]. Die internen Schichten der eingesetzten Künstlichen Neuronalen Netze (KNN) erkennen zwar komplexe Muster und Zusammenhänge, ihre „Schlussfolgerungen" sind jedoch in der Regel für Menschen nicht unmittelbar nachvollziehbar oder gar wörtlich und inhaltlich ablesbar. Für den Menschen bleibt daher unklar, welche konkreten Wahrnehmungen und Bewertungen zu einer bestimmten Fahrentscheidung geführt haben[7].

An diesem Punkt setzt Explainable AI (XAI) an. Sie soll dazu dienen, die für eine Entscheidung relevanten Schritte und Zusammenhänge innerhalb eines KI-Systems in menschenlesbare Erklärungen zu übersetzen. Im automobilen Kontext sollen hierdurch insbesondere die Vorhersehbarkeit des automatisierten Fahrsystems gestärkt und Unsicherheiten aufseiten der Fahrgäste reduziert werden. Grundsätzlich lassen sich zwei zentrale Strategien unterscheiden:

1. Post-hoc-Erklärbarkeit:

Bei diesem Ansatz wird die Entscheidung eines bereits trainierten, komplexen KI-Modells nachträglich analysiert, ohne die ursprüngliche Architektur grundlegend zu verändern. Verfahren wie *Saliency Maps* markieren beispielsweise diejenigen Bereiche eines Kamerabildes, die für die Entscheidung des Systems besonders relevant waren. Auf diese Weise lässt sich im Nachhinein rekonstruieren, ob etwa ein*e Fußgänger*in, ein Hindernis oder eine Fahrbahnmarkierung den Ausschlag für ein Brems- oder Ausweichmanöver gegeben hat. Die Erklärung bildet den Entscheidungsprozess dabei nicht unmittelbar ab, sondern erfolgt im Nachhinein durch Interpretation.[8]

2. Ante-hoc-Erklärbarkeit:

Dagegen wird bei diesen Modellen die Erklärbarkeit bereits in die Architektur und den Entscheidungsprozess des Systems integriert. Ein Beispiel hierfür sind sogenannte Concept-Bottleneck-Models. Sie übersetzen hochdimensionale Sensordaten zunächst in explizit definierte und menschenlesbare Informationen wie „Fußgänger*in voraus", „Sicherheitsabstand unterschritten" oder „eingeschränkte Sichtweite". Erst auf Grundlage dieser Zwischenzustände wird die eigentliche Fahrentscheidung getroffen. Da die Begründung damit strukturell Teil des Entscheidungsprozesses ist, kann das System einen kausalen Zusammenhang zwischen Wahrnehmung und Reaktion herstellen, der dem Menschen nachvollziehbar ist.[9]

Während diese XAI-Ansätze zeigen, wie Erklärungen technisch erzeugt werden können, stellt sich anschließend die Frage, welche Informationen einem Fahrgast in einer konkreten Fahrsituation zur Verfügung gestellt werden müssen. Einen geeigneten Ordnungsrahmen bietet hierfür das „Situation-Awareness-Based Agent Transparency Model", kurz SAT-Modell. Es unterscheidet drei aufeinander aufbauende Transparenzstufen[10]:

[6] s. Li 2026, S. 253
[7] s. Esser et al. 2020, S. 1
[8] s. Kuznietsov et al. 2024, S. 19346
[9] s. Kuznietsov et al. 2024, S. 19346
[10] s. Chen et al. 2018, S. 261f.

Transparenzstufe 1 – Systemstatus:
Die erste Stufe vermittelt grundlegende Informationen über den aktuellen Zustand und die ausgeführte Handlung des Systems. Dazu zählen beispielsweise der aktive Fahrmodus, die erkannte Verkehrssituation oder ein eingeleiteter Bremsvorgang. Da nur das aktuelle, situative Systemverhalten abgebildet wird, sind hierfür noch keine umfassenden XAI-Methoden erforderlich.

Transparenzstufe 2 – Handlungslogik:
Die zweite Stufe ergänzt den Systemstatus um die diesem zugrunde liegende Begründung. Sie erklärt, warum das Fahrzeug eine bestimmte Reaktion ausführt. Post-hoc-Methoden können hierzu nachträglich diejenigen Wahrnehmungsmerkmale hervorheben, die für eine Entscheidung besonders relevant waren. Bei Ante-hoc-Modellen ist die Begründung bereits als nachvollziehbarer Zwischenschritt Teil des Entscheidungsprozesses.

Transparenzstufe 3 – Zukunftsprojektion:
In der dritten Stufe wird die Erklärung um die erwartete weitere Entwicklung ergänzt. Kommuniziert werden geplante Folgeschritte und deren voraussichtliche Konsequenzen, beispielsweise eine bevorstehende Spuränderung, die geplante Trajektorie oder eine zu erwartende Verzögerung. Ante-hoc-Modelle können diese Zusammenhänge bereits während der Planung nachvollziehbar machen. Post-hoc-Methoden analysieren dagegen die erzeugten Trajektorien und visualisieren die prognostizierte Dynamik des eigenen Fahrzeugs sowie anderer Verkehrsteilnehmer*innen.

XAI schafft somit die technische Grundlage, um interne Systementscheidungen in für den Menschen nachvollziehbare Informationen zu übersetzen. Das SAT-Modell strukturiert diese Informationen anschließend nach ihrer inhaltlichen Tiefe – vom aktuellen Systemstatus über die Begründung einer Handlung bis hin zur Projektion zukünftiger Entwicklungen. Damit ist jedoch noch nicht geklärt, auf welche Weise diese Erklärungen den Fahrgast tatsächlich erreichen. Diese Übersetzungsaufgabe übernimmt das Human-Machine-Interface.

Zukunftsfähige Human-Machine-Interfaces

Die im SAT-Modell beschriebenen Transparenzstufen definieren, welche Informationen ein autonomes Fahrzeug bereitstellen muss. Damit die Informationen für den Fahrgast tatsächlich verständlich und handlungsrelevant werden, müssen sie jedoch in geeignete Kommunikationsformen übersetzt werden. Diese Aufgabe übernimmt das Human-Machine-Interface. Es bildet die zentrale Schnittstelle zwischen dem automatisierten Fahrsystem und seinen Fahrgästen und macht die Wahrnehmungen, Entscheidungen und Handlungsabsichten des Systems visuell, akustisch oder haptisch erfahrbar[11].

Die Gestaltung eines solchen Interface beschränkt sich daher nicht auf die bloße Anzeige des aktuellen Systemstatus: Vielmehr muss für jede Transparenzstufe bestimmt werden, welche Informationen in einer konkreten Fahrsituation relevant sind, über welchen Wahrnehmungskanal sie vermittelt werden und in welcher Darstellungsform sie den Fahrgast am wirksamsten erreichen[12]. Während grundlegende Statusinformationen beispielsweise dauerhaft visuell angezeigt werden können, erfordern sicherheitskritische Hinweise eine unmittelbar wahrnehmbare und eindeutig zu interpretierende Form.

[11] s. Mandujano-Granillo et al. 2024, S. 121638–121642
[12] s. Hong et al. 2025, S. 1350; S. 1360f.

Wenn die Informationsübertragung nur einen menschlichen Wahrnehmungssinn anspricht, kann dies vor allem in schwierigen oder zeitkritischen Fahrsituationen zu Problemen führen. Visuelle Hinweise können übersehen werden, wenn der Blick des Fahrgastes nicht auf die Fahrbahn oder das Display gerichtet ist. Akustische Signale sind zwar unabhängig von der Blickrichtung wahrnehmbar, sind jedoch nicht eindeutig einem Objekt oder einem Bereich zuzuordnen. Haptische Reize , wie ein Vibration des Lenkrads, könnten die Nutzer*innen zwar unmittelbar erreichen, würden sich aber vor allem für Warnungen oder Handlungsaufforderungen eignen. Die Auswahl des Kommunikationskanals muss daher kontextabhängig erfolgen.[13]

Zukunftsfähige HMI-Architekturen kombinieren deshalb visuelle, akustische und haptische Elemente zu einer multimodalen Kommunikation. Durch die koordinierte Nutzung mehrerer Wahrnehmungskanäle können Informationen schneller von den Fahrgästen wahrgenommen und mentaler Stress reduziert werden. Voraussetzung hierfür ist jedoch, dass die einzelnen Signale inhaltlich sowie räumlich und zeitlich aufeinander abgestimmt sind. Widersprüchliche, verzögerte oder sich überlagernde Reize können irritieren und das Vertrauen in das System schmälern[14]. Entscheidend ist daher nicht die größtmögliche Anzahl an Informationen, sondern deren situationsgerechte Auswahl, Priorisierung und Darstellung.

Die Anforderungen an das HMI lassen sich direkt aus den drei Transparenzstufen des SAT-Modells ableiten: Auf der ersten Stufe muss das Interface erkennbar machen, in welchem Betriebszustand sich das Fahrzeug befindet und welche Handlung es aktuell ausführt. Die zweite Stufe erfordert zusätzlich eine verständliche Begründung, etwa durch die Kennzeichnung des erkannten Hindernisses oder eine kurze sprachliche Erklärung des Bremsmanövers. Auf der dritten Stufe muss das Fahrzeug darüber hinaus seine nächste Handlung mitteilen, beispielsweise durch die Visualisierung einer geplanten, abweichenden Fahrbewegung oder die Ankündigung eines bevorstehenden Spurwechsels.

Ein technologisch vielversprechender Ansatz zur Umsetzung dieser Anforderungen sind Augmented-Reality-Head-up-Displays. Durch sie haben die Fahrgäste die KI-generierten Informationen direkt vor Augen und somit eingebettet in den räumlichen Kontext der Fahrbahn und des Geschehens. Erkannte Verkehrsteilnehmer*innen, Hindernisse, Fahrwege oder Systemgrenzen können direkt an ihrer tatsächlichen Position in der Umgebung hervorgehoben werden[15]. Dadurch wird nicht nur direkt erkennbar, dass das Fahrzeug gleich reagieren wird, sondern auch, worauf es reagiert. Ein aktuelles Implementierungsbeispiel für Augmented-Reality-Head-up-Displays liefert die im Jahr 2025 vorgestellte Kooperation zwischen *Xpeng* und *Huawei*[16]. Das System „Chasing Light Panorama“ blendet Navigations- und Assistenzinformationen situationsabhängig in das reale Fahrgeschehen ein. Hierzu werden Kamera- und Radardaten zusammengeführt, um erkannte Objekte und Fahrinformationen lagegerecht in das Sichtfeld der Umgebung einzubetten[17]. Auf diese Weise wird die Frontscheibe von einer reinen Anzeigefläche zu einer kontextsensitiven Kommunikationsschnittstelle zwischen Fahrzeug, Verkehrssituation und Fahrgast.

---

[13] s. Arabian et al. 2025, S. 2; S. 15

[14] s. Ma et al. 2022, S. 10

[15] s. Mandujano-Granillo et al. 2024, S. 121643

[16] https://www.huawei.com/de/news/2021/huawei-ar-hud-iaa-mobility-2021 (Link zuletzt geprüft: 08.10.2026)

[17] s. Bork 2025

Zukunftsfähige HMIs machen die von XAI-Systemen erzeugten Erklärungen somit erstmals für den Menschen wahrnehmbar und interpretierbar. Ihre konkrete Ausgestaltung hängt jedoch wesentlich davon ab, welche Rolle der Mensch im jeweiligen Automatisierungsszenario einnimmt. Während das Interface auf niedrigeren Automatisierungsstufen vor allem die Überwachung und gegebenenfalls die Rückübernahme der Fahraufgabe unterstützen muss, rücken auf höheren Stufen die subjektive Sicherheit, die Nachvollziehbarkeit autonomer Manöver und die Qualität des Fahrerlebnisses in den Mittelpunkt. Wie sich diese unterschiedlichen Anforderungen praktisch umsetzen lassen, zeigen die nachfolgenden Implementierungsbeispiele entlang der SAE-Level.

Implementierungsbeispiele entlang der SAE-Level

Wie Explainable AI und Human-Machine-Interfaces konkret ausgestaltet werden müssen, hängt wesentlich vom Automatisierungsgrad des Fahrzeugs ab. Die Klassifizierung der *Society of Automotive Engineers* unterscheidet sechs Stufen von SAE-Level 0 bis 5[18]. Während der Mensch bis einschließlich Level 2 die Fahraufgabe kontinuierlich überwachen muss, übernimmt das automatisierte Fahrsystem ab Level 3 zunehmend selbstständig die Fahrzeugführung[19]. Damit verändert sich zugleich die Funktion der Kommunikation: Sie reicht von der sicheren Rückübertragung der Fahraufgabe über die Erklärung autonomer Manöver bis zur vorausschauenden Gestaltung des Fahrerlebnisses. Aktuelle Serienlösungen und Forschungsprojekte veranschaulichen, wie diese Anforderungen technisch umgesetzt werden können.

Auf SAE-Level 3 steht die sichere Übergabe zwischen Mensch und System im Mittelpunkt.

Mit dem Übergang zur bedingten Automatisierung kann die fahrende Person unter den jeweils zugelassenen Betriebsbedingungen zeitweise das Fahren dem Auto überlassen. Kommt das System jedoch an seine funktionalen oder räumlichen Grenzen, muss der*die Fahrer*in nach einer Übernahmeaufforderung, dem sogenannten Take-over-Request (TOR), die Fahrzeugführung wieder komplett selbst übernehmen[20]. Diese Übergabesituation ist eine besondere Herausforderung: Da der*die Fahrer*in das Verkehrsgeschehen zuvor nicht lückenlos beobachtet hat, muss er*sie innerhalb kurzer Zeit ein ausreichendes Situationsbewusstsein aufbauen, um das Fahren spontan, aber sicher zu übernehmen[21]. Das HMI darf daher nicht nur zur Übernahme auffordern, sondern muss die Aufmerksamkeit des*der Fahrer*in gezielt zurück auf das Verkehrsgeschehen lenken und die Dringlichkeit der Situation verdeutlichen.

Wie stark die Wahl der Kommunikationskanäle die Reaktionsfähigkeit der Menschen beeinflusst, zeigt eine Simulatorstudie des *Würzburger Instituts für Verkehrswissenschaften* (WIVW). Bei einer ausschließlich visuellen Übernahmeaufforderung benötigten die Versuchspersonen unter Ablenkungsbedingungen im Durchschnitt 6,2 Sekunden, um das Lenkrad wieder zu übernehmen. Durch die Kombination visueller und akustischer Warnreize

[18] s. Mandujano-Granillo et al. 2024, S. 121638
[19] s. Mandujano-Granillo et al. 2024, S. 121638
[20] s. Petermeijer et al. 2017, S. 204
[21] s. Petermeijer et al. 2017, S. 205f.

konnte die Reaktionszeit verkürzt und die Qualität der Übernahme verbessert werden[22]. Dieses Beispiel unterstreicht die Bedeutung multimodaler Interfaces insbesondere in zeitkritischen Situationen. Eine entsprechende Umsetzung erfolgte bereits im *Mercedes-Benz Drive Pilot*. Der aktive Systemzustand wird unter anderem durch eine türkisfarbene Beleuchtung am Lenkrad angezeigt. Wird eine Übernahme erforderlich, erfolgen zunächst eine kombinierte visuelle und akustische Warnung. Reagiert der*die Fahrer*in nicht, wird die Aufforderung durch einen haptischen Lenkradimpuls verstärkt[23]. Die Kommunikation passt sich damit der Dringlichkeit der Situation an: Je dringender die Reaktion nötig wird, desto intensiver erfolgen die Meldungen und desto mehr Kommunikationskanäle werden verwendet.

Auf SAE-Level 4 verlagert sich der Schwerpunkt von der Übernahmefähigkeit zur Nachvollziehbarkeit autonomer Manöver.

Innerhalb der festgelegten Betriebsumgebung, der Operational-Design-Domain (ODD), übernimmt das Fahrzeug das Fahren vollständig. Eine „menschliche Notübernahme“ ist dort nicht mehr vorgesehen. Kommt das System an seine Grenzen oder tritt eine Störung auf, muss das Fahrzeug selbstständig das Risiko verringern, beispielsweise durch kontrolliertes Anhalten[24]. Das HMI dient auf dieser Stufe ergo nicht mehr vor allem dazu, die unmittelbare Übernahme der Fahrzeugführung durch den Menschen vorzubereiten. Stattdessen soll es das subjektive Sicherheitsgefühl der Fahrgäste stärken und die Fahrmanöver nachvollziehbar gestalten. Wenn das Fahrzeug unerwartet bremst oder ausweicht, kann dies ohne nachvollziehbare Erklärung bei den Fahrgästen ein Gefühl des Kontrollverlusts hervorrufen. Selbst ein objektiv korrektes Fahrverhalten kann dadurch als unsicher oder unberechenbar wahrgenommen werden[25]. Erklärungsfähige Systeme müssen deshalb nicht nur mitteilen, was abweichend geschieht, sondern auch den Grund und die weitere Planung abbilden können. Ein solcher Ansatz wird durch das im Juni 2026 vorgestellte *Alpamayo 2 Super Framework* von NVIDIA verfolgt. Das für autonome Robotaxis entwickelte Modell erzeugt die Fahrentscheidung und ihre sprachliche Begründung simultan. Die Erklärung wird den Fahrgästen nicht erst im Nachhinein aus der bereits getroffenen Entscheidung heraus erklärt, sondern unmittelbar mit ihr verknüpft[26].

Auf SAE-Level 5 wird die Kommunikation schließlich zu einem integralen Bestandteil des Reiseerlebnisses.

Auf der höchsten Automatisierungsstufe sind Aktivitäten des Menschen nicht mehr vorgesehen; auch bleibt das Fahrerlebnis nicht mehr auf eine festgelegte Betriebsumgebung beschränkt. Das Fahrzeug vollzieht damit perspektivisch endgültig den Wandel von einem klassischen Verkehrsmittel hin zu einem mobilen Lebens-, Arbeits- und Erholungsraum[27]. Da sich die Fahrgäste hier dauerhaft anderen Tätigkeiten wie Arbeiten, Lesen oder Schlafen zuwenden können, entstehen neue Anforderungen an die Kommunikation des Fahrzeugverhaltens. Vor allem im Mittelpunkt steht hier die Vermeidung der

[22] s. Naujoks et al. 2021, S. 431–438
[23] s. Mercedes-Benz USA 2021, S. 7f.
[24] s. SAE J3016
[25] s. Hong et al. 2025, S. 1350f.
[26] s. Labrie 2026
[27] s. SAE J3016

als Reisekrankheit bekannten Beschwerden beim Autofahren. Wenn der Mensch nicht auf die Fahrbahn sieht, stimmen die optisch wahrgenommene Ruhe und die durch den Gleichgewichtssinn registrierte Fahrzeugbewegung nicht mehr überein. Beschleunigen, Bremsen oder abruptes Umlenken können dadurch Reisekrankheit auslösen[28]. Prädiktive HMI-Konzepte berücksichtigen dies bereits vor der eigentlichen Fahrbewegungsänderung: Sie informieren die Fahrgäste nicht nur über die aktuelle Fahrsituation oder das anstehende Manöver, sondern kündigen die bevorstehende Handlung frühzeitig an. Der Fahrgast erhält dadurch die Möglichkeit, sich bewusst oder auch unbewusst auf eine Kurve, eine Beschleunigung oder einen Bremsvorgang einzustellen. Das Deutsche Zentrum für Luft- und Raumfahrt untersucht hierzu unter anderem den Einsatz unterschwelliger HMI-Reize. Kurz vor einem bevorstehenden, für den Fahrgast überraschenden Fahrmanöver können beispielsweise dezente Lichtbewegungen im Fahrzeuginnenraum oder gezielte akustische Signale die geplante Bewegungsrichtung ankündigen. Das menschliche Gehirn ist dadurch auf die nachfolgende Fahrzeugbewegung besser vorbereitet, wodurch der Wechsel im Fahrablauf nicht mehr als so abrupt wahrgenommen wird und das Einsetzen der Reisekrankheit verhindert werden kann.[29]

Die genannten Implementierungsbeispiele verdeutlichen, dass sich die Aufgabe der Fahrzeugkommunikation entlang der SAE-Level grundlegend verändert: Auf Level 3 muss sie vor allem eine schnelle und sichere Rückübernahme der Fahraufgabe durch den Menschen ermöglichen. Auf Level 4 steht die verständliche Erklärung autonomer Entscheidungen des Fahrzeugs im Vordergrund. Auf Level 5 wird die vorausschauende Kommunikation zu einem zentralen Bestandteil von Komfort- und Fahrerlebnis der passiv fahrenden Fahrgäste. Mit zunehmender Automatisierung steigt damit nicht nur die technische Verantwortung des Systems, sondern auch die Bedeutung einer Kommunikation, die dessen Wahrnehmungen, Entscheidungen und Absichten nachvollziehbar vermittelt.

---

[28] s. Metzulat et al. 2024, S. 234f.

[29] s. Begli 2021

Regulatorischer Rahmen und juristische Implikationen

Die zunehmende Automatisierung des Fahrens hat nicht nur technische und gestalterische Konsequenzen. Je stärker das Fahrzeug die Fahraufgabe selbstständig übernimmt, desto mehr verlagern sich Verantwortung, Kontrollpflichten und Haftungsrisiken von den Fahrer*innen auf das automatisierte System, die Halter*innen und den Hersteller. Transparenz entwickelt sich damit von einem vertrauensbildenden Gestaltungsmerkmal zu einer wesentlichen Voraussetzung für Zulassung, Nachweisführung und rechtliche Verantwortungszuordnung.

Bis einschließlich SAE-Level 2 bleibt der Mensch dauerhaft für die Überwachung des Verkehrsgeschehens und die sichere Fahrzeugführung verantwortlich. Die Verschuldenshaftung des*der Fahrer*in bildet daher weiterhin das dominierende rechtliche Prinzip[30].

Mit dem Übergang zu SAE-Level 3 verändert sich diese Verantwortungsverteilung. Innerhalb der freigegebenen Betriebsbedingungen übernimmt das automatisierte Fahrsystem die Fahraufgabe, während sich der*die Fahrer*in zeitweise vom Verkehrsgeschehen abwenden kann. In diesen automatisierten Fahrphasen tritt die klassische Fahrzeugführungsverantwortung zurück, während die technische Funktionsfähigkeit des Systems und damit die Produktverantwortung des Herstellers stärker in den Mittelpunkt rücken[31]. Besonders sensibel ist dabei die Schnittstelle zwischen automatisierter und menschlicher Fahrzeugführung. Kommt das System an seine funktionalen oder räumlichen Grenzen, muss es den*die Fahrer*in rechtzeitig, eindeutig und situationsgerecht zur Übernahme des Steuers auffordern. Dieser Take-over-Request darf nicht auf ein einzelnes Warnsignal beschränkt sein, sondern muss es dem*der Fahrer*in ermöglichen, innerhalb der meist knappen, verfügbaren Zeit die Situation und die Dringlichkeit zu erfassen und die Fahrverantwortung sicher übernehmen zu können. Bleibt eine Reaktion des Menschen aus, muss das Fahrzeug selbstständig ein möglichst risikoarmes Manöver einleiten. Für die rechtliche Bewertung im Nachgang ist daher nicht allein entscheidend, ob eine Übernahmeaufforderung an den Menschen erfolgte. Ebenso relevant sind Zeitpunkt, Verständlichkeit und technische Zuverlässigkeit sowie die Frage, ob der*die Fahrer*in tatsächlich in die Lage versetzt wurde, die Kontrolle sicher zurückzugewinnen. Fehlerhafte oder verspätete Warnungen, unklar kommunizierte Systemgrenzen und eine unzureichende Übergabe an den*die Fahrer*in können erhebliche Haftungsrisiken für den Hersteller begründen.[32]

Auf SAE-Level 4 entfällt die Rückfallebene auf menschliche Übernahme innerhalb der definierten Operational-Design-Domain weitgehend. Das Fahrzeug muss auch bei Systemgrenzen oder Störungen selbstständig einen sicheren beziehungsweise möglichst risikoarmen Zustand herstellen. Der Mensch übernimmt hier nicht mehr unmittelbar die Fahraufgabe; abhängig vom Anwendungsszenario kann jedoch eine technische Aufsicht vorgesehen sein.[33]

Auf SAE-Level 5 ist schließlich auch die Beschränkung auf eine festgelegte Betriebsumgebung nicht mehr gegeben. Das automatisierte System trägt hier die Verantwortung für die Fahrzeugführung unter sämtlichen Verkehrs- und Umgebungsbedin-

[30] s. Mercedes-Benz Group AG 2025
[31] s. Mercedes-Benz Group AG 2025
[32] s. Park 2024, S. 41
[33] s. SAE J3016

gungen. Mit dieser Entwicklung verliert der*die menschliche Fahrer*in als unmittelbares Haftungssubjekt der Fahraufgabe endgültig an Bedeutung. Stattdessen gewinnen die verschuldensunabhängige Halterhaftung sowie die Produkt- und Herstellerhaftung an Gewicht. Gleichzeitig steigen die Anforderungen an die technische Dokumentation. Im Schadensfall muss nachvollziehbar sein, welche Umweltinformationen das Fahrzeug erfasst hat, wie es die Situation bewertet hat, welche Handlungsoptionen berücksichtigt wurden und warum schließlich eine bestimmte Entscheidung getroffen wurde. Erklärbarkeit wird damit nicht nur für den Fahrgast, sondern auch für Behörden, Sachverständige, Versicherungen und Gerichte relevant.[34]

Diese Entwicklung wird durch die europäische KI-Regulierung zusätzlich verstärkt. KI-Systeme, die als sicherheitsrelevante Bestandteile automatisierter Fahrzeuge eingesetzt werden, können abhängig von ihrer konkreten Ausgestaltung und regulatorischen Einordnung den Anforderungen an Hochrisiko-KI-Systeme unterliegen[35]. Damit verbinden sich unter anderem Pflichten zur Risikosteuerung, technischen Dokumentation, menschlichen Aufsicht und Transparenz.

Artikel 13 des *EU AI Act* verlangt, dass Hochrisiko-KI-Systeme so transparent gestaltet werden, dass ihre Funktionsweise für die vorgesehenen Nutzer*innen hinreichend nachvollziehbar ist.[36]Für autonome Fahrzeuge entsteht daraus eine unmittelbare Gestaltungsanforderung an das HMI: Das System muss seinen Betriebszustand, seine Grenzen und die für die Nutzer*innen relevanten Handlungen so kommunizieren, dass eine sachgerechte und sichere Nutzung möglich wird.

Ergänzend fordert Artikel 12 des *EU AI Act* die automatische Aufzeichnung relevanter Systemereignisse.[37] Eine solche Protokollierung soll ermöglichen, Fehlfunktionen, kritische Situationen und Unfälle nachträglich zu rekonstruieren und die Einhaltung regulatorischer Anforderungen zu überprüfen. Die im Fahrzeug angezeigte Erklärung und die intern dokumentierte Entscheidungsgrundlage müssen dabei übereinstimmen. Gemäß der Verordnung 2024/1689 des Europäischen Parlaments und des Europäischen Rates reicht eine nach außen nachvollziehbar wirkende Begründung nicht aus, wenn sie dem tatsächlichen Entscheidungsprozess des Systems nicht voll entspricht[38].

Transparente Fahrzeugkommunikation erfüllt damit eine doppelte Funktion: Gegenüber dem Fahrgast schafft sie Orientierung, Berechenbarkeit und Vertrauen. Gegenüber Aufsichtsbehörden und der Rechtsprechung ermöglicht sie die Prüfung, Rekonstruktion und rechtliche Bewertung autonomer Systementscheidungen. Erklärbarkeit wird dadurch zu einem Bindeglied zwischen technischer Leistungsfähigkeit, gesellschaftlicher Akzeptanz und regulatorischer Zulassungsfähigkeit.

---

[34] s. Park 2024, S. 13; S. 22–25

[35] s. Tanenblatt et al. 2025, S. 26

[36] s. EU Artificial Intelligence Act, Artikel 13

[37] s. EU Artificial Intelligence Act, Artikel 12

[28] s. EU 2024/1689

4 Fazit und Ausblick

Der Weg zum hochautomatisierten Fahren entscheidet sich nicht allein auf der Straße, sondern auch in den Köpfen der Menschen. Ängste, Bedenken und Unsicherheiten der Fahrgäste können die Akzeptanz autonomer Fahrzeuge erheblich beeinträchtigen und damit die Transformation hin zu einer neuen Form der Mobilität insbesondere in Deutschland ausbremsen. Neben der technischen Sicherheit müssen deshalb auch Transparenz und Nachvollziehbarkeit der vom System getroffenen Entscheidungen gewährleistet werden. Mit Explainable-AI-Ansätzen, transparenzorientierten Modellen wie dem SAT-Modell und multimodalen Human-Machine-Interfaces bestehen hierfür sowohl in der Forschung als auch in ersten praktischen Anwendungen vielversprechende Lösungsansätze. Sie ermöglichen es, den aktuellen Systemzustand, die Gründe für eine Fahrentscheidung und die erwarteten nächsten Handlungsschritte verständlich aufzubereiten. Dadurch lässt sich die bislang als „Black Box“ wahrgenommene Entscheidungslogik zumindest teilweise öffnen und das Fahrzeugverhalten für Fahrgäste berechenbarer gestalten.

Damit diese Ansätze über einzelne Anwendungsfälle hinaus wirksam werden, sind jedoch weiterführende Untersuchungen erforderlich. Die Forschung muss klären, welche Informationen Fahrgäste in unterschiedlichen Situationen tatsächlich benötigen und wann visuelle, akustische oder haptische Reize eingesetzt werden sollten. Ebenso ist zu untersuchen, wie diese Reize sinnvoll kombiniert und dosiert werden können, damit sie Orientierung schaffen, ohne die Fahrgäste zu überfordern oder neue Unsicherheit auszulösen. Dabei sind sowohl die unterschiedlichen Anforderungen der SAE-Level 3, 4 und 5 als auch verschiedene Verkehrsteilnehmer*innen, Ablenkungszustände und Fahrsituationen zu berücksichtigen.

Auf dieser Grundlage müssen anschließend auch die regulatorischen Anforderungen weiter konkretisiert werden. Für jedes Automatisierungslevel ist eindeutig festzulegen, welche Informationen das Fahrzeug in welcher Situation, zu welchem Zeitpunkt und über welchen Kommunikationskanal vermitteln muss.

Literaturverzeichnis

Ali, F. N.; Latif, O. A.: Explainable AI in Autonomous Driving: A Comparative Study of Post-Hoc Explanations for YOLOv8. In: IEEE Access 14 (2026), S. 66358–66371. DOI: 10.1109/ACCESS.2026.3689120

Arabian, A.; Goodridge, C. M.; Merat, N.: Asynchrony of directional auditory warnings and visual information in hands-on SAE level 2 automated vehicles: Analysis of driver responses. In: PLOS ONE 20 (2025) 11, Art.-nr. e0336941. DOI: 10.1371/journal.pone.0336941

Begli, J.: [Pressemitteilung] Projekt AutoAkzept: „Wohlfühlen, Arbeiten und Ausruhen im automatisierten Fahrzeug". DLR online, 24.02.2021. https://www.dlr.de/de/aktuelles/nachrichten/2021/01/20210224_projekt-autoakzept (Link zuletzt geprüft: 23.07.2026)

Bork, H.: Mit Huawei: Xpeng macht Windschutzscheibe zum AR-Bildschirm. Automobil Industrie online, 01.07.2025. https://www.automobil-industrie.vogel.de/xpeng-huawei-augmented-reality-head-up-display-a-5ce566f945e0e1cd8ce428f46652e9d0/ (Link zuletzt geprüft: 23.07.2026)

Chen, J. Y. C.; Lakhmani, S. G.; Stowers, K.; Selkowitz, A. R.; Wright, J. L.; Barnes, M.: Situation awareness-based agent transparency and human-autonomy teaming effectiveness. In: Theoretical Issues in Ergonomics Science 19 (2018) 3, S. 259–282. DOI: 10.1080/1463922X.2017.1315750

Deng, Y.; He, C.: A User-driven Design Framework for Robotaxi. [Submitted on 22 Feb 2026 (v1), last revised 3 Jul 2026 (this version, v3)] DOI: 10.48550/arXiv.2602.19107

Esser, P.; Rombach, R.; Ommer, B.: A Disentangling Invertible Interpretation Network for Explaining Latent Representations. Arxiv online, 27.04.2020. DOI: 10.48550/arXiv.2004.13166

Europäische Union (Hrsg.): Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024 zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz und zur Änderung der Verordnungen (EG) Nr. 300/2008, (EU) Nr. 167/2013, (EU) Nr. 168/2013, (EU) 2018/858, (EU) 2018/1139 und (EU) 2019/2144 sowie der Richtlinien 2014/90/EU, (EU) 2016/797 und (EU) 2020/1828 (Verordnung über künstliche Intelligenz), 12.07.2024. https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=OJ:L_202401689 (Link zuletzt geprüft: 12.08.2026)

Hong, J.; Kim, S.; Lee, S.: Enhancing passenger-vehicle interaction through multimodal explanation for unexpected behaviors of fully autonomous driving in non-driving-related tasks. In: Transportation Research Part F: Traffic Psychology and Behaviour 109 (2025), S. 1350–1364. DOI: 10.1016/j.trf.2025.01.045

Kusano, K. D.; Scanlon, J. M.; Chen, Y.-H.; McMurry, T. L.; Gode, T.; Victor, T.: Comparison of Waymo Rider-Only Crash Rates by Crash Type to Human Benchmarks at 56.7 Million Miles. arXiv online, 02.05.2025. DOI: 10.48550/arXiv.2505.01515

Kuznietsov, A.; Gyevnar, B.; Wang, C.; Peters, S.; Albrecht, S. V.: Explainable AI for Safe and Trustworthy Autonomous Driving: A Systematic Review. In: IEEE Transactions on Intelligent Transportation Systems 25 (2024) 12, S. 19432 19364. DOI: 10.1109/TITS.2024.3474469

Labrie, M.: [Pressemitteilung]NVIDIA Launches Alpamayo 2 Super Open Reasoning Model for Robotaxis. NVIDIA online, 31.05.2026. https://nvidianews.nvidia.com/news/nvidia-alpamayo-2-super-robotaxis (Link zuletzt geprüft: 23.07.2026)

Li, Y.: Using Explainable Artificial Intelligence (X-AI) To Address The "Black-Box" Dilemma in Autonomous Driving. In: Mathematical Modeling and Algorithm Application 9 (2026) 1, S. 252–259. DOI: 10.54097/fphdfy81

Luo, X.; Ding, F.; Chen, R.; Panda, R.; Loo, J.; Zhang, S.: „What's Happening"- A Human-centered Multimodal Interpreter Explaining the Actions of Autonomous Vehicles. [Submitted on 9 Jan 2025 (v1), last revised 12 Jan 2025 (this version, v2)] Arxiv online, 12.01.2025. DOI: 10.48550/arXiv.2501.05322

Ma, J.; Li, J.; Huang, H.: Evaluation of Multimodal and Multi-Staged Alerting Strategies for Forward Collision Warning Systems. In: Sensors 22 (2022) 3, Art.-nr. 1189, 12 S. DOI: 10.3390/s22031189

Mandujano-Granillo, J. A.; Candela-Leal, M. O.; Ortiz-Vazquez, J. J.; Ramirez-Moreno, M. A.; Tudon-Martinez, J. C.; Felix-Herran, L. C.; Galvan-Galvan, A.; Lozoya-Santos, J. D. J.: Human–Machine Interfaces: A Review for Autonomous Electric Vehicles. In: IEEE Access 12 (2024), S. 121635–121658. DOI: 10.1109/ACCESS.2024.3450439

Mercedes-Benz Group AG (Hrsg.): Automatisiertes und autonomes Fahren. Recht und Ethik. Mercedes-Benz online, November 2025. https://group.mercedes-benz.com/innovation/case/autonomous/rechtlicher-rahmen.html (Link zuletzt geprüft: 23.07.2026)

Mercedes-Benz USA (Hrsg.): DRIVE PILOT First Responder Interaction Plan. Mercedes-Benz USA online, April 2021. https://www.mbusa.com/content/dam/mb-nafta/us/owners/drive-pilot/DRIVE_PILOT_First_Responder_Interaction_Plan.pdf (Link zuletzt geprüft: 23.07.2026)

Metzulat, M.; Metz, B.; Landau, A.; Neukum, A.; Kunde, W.: Does the visual input matter? Influence of non-driving related tasks on car sickness in an open road setting. In: Transportation Research Part F: Traffic Psychology and Behaviour 104 (2024), S. 234–248. DOI: 10.1016/j.trf.2024.06.002

Naujoks, F.; Mai, C.; Neukum, A.: The Effect of Urgency of Take-Over Requests During Highly Automated Driving Under Distraction Conditions. In: Advances in Human Aspects of Transportation: Part I. AHFE International - Applied Human Factors and Ergonomics International 15 (2021), S. 2099–2106. DOI: 10.54941/ahfe100646

Park, J.: Strafrechtliche Verantwortlichkeit im Zusammenhang mit autonomen Fahrzeugen: Problematik und Lösungsansätze. Berlin, Freie Univ., Diss., 2024. https://refubium.fu-berlin.de/handle/fub188/45710 (Link zuletzt geprüft: 23.07.2026)

Petermeijer, S.; Bazilinskyy, P.; Bengler, K.; de Winter, J.: Take-over again: Investigating multimodal and directional TORs to get the driver back into the loop. In: Applied Ergonomics 62 (2017), S. 204–215. DOI: 10.1016/j.apergo.2017.02.023

SAE J3016: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. SAE International/SAE Mobilus, 30.04.2021. https://saemobilus.sae.org/standards/j3016_202104-taxonomy-definitions-terms-related-driving-automation-systems-road-motor-vehicles (Link zuletzt geprüft: 23.07.2026)

Tanenblatt, E.; Stockburger, P.; Boothe, W.; Benzadon, J.; Daubert, T.; Denham J.; Cooper, J.: 2025 Global Guide to Autonomous Vehicles. Dentons, London [u. a.] 22.05.2025. https://www.dentons.com/en/insights/guides-reports-and-whitepapers/2025/may/14/-/media/519AF07BF6574A568759FAC89E64BD8B.ashx (Link zuletzt geprüft: 23.07.2026)

Abkürzungsverzeichnis

HMI: *Human Machine Interface*

KNN: Künstliche Neuronale Netze

SAT-Modell: *Situation Awareness Based Agent Transparency Model*

XAI: *Explainable AI*

DiSer Hub

Zuwendungsgeber:

Gefördert durch:

Bundesministerium für Wirtschaft und Energie

Förderkennzeichen: 16THB0004A
Laufzeit: 01 09 2022 – 31 12 2026
Projektträger:

VDI | VDE | IT

VDI / VDE Innovation + Technik GmbH

5 Partner. 5 Standorte. 1 Netzwerk.

fir an der RWTH Aachen

Fraunhofer AUTOMOBIL

ifa Institut für Automobilwirtschaft

SICP Software Innovation Campus Paderborn
HEINZ NIXDORF INSTITUT
UNIVERSITÄT PADERBORN

TECHNISCHE UNIVERSITÄT CHEMNITZ

TUCed
AN-INSTITUT FÜR TRANSFER UND WEITERBILDUNG

diserhub.de