Lernende Systeme

DIE PLATTFORM FÜR KÜNSTLICHE INTELLIGENZ

Vertrauen in KI-basierte Mobilität

Technologische und ethische Aspekte

WHITEPAPER

Bahlmann, C., Felix, R., Hahn, A. et al.
AG Mobilität und intelligente Verkehrssysteme
AG IT-Sicherheit, Privacy, Recht und Ethik

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

acatech
DEUTSCHE AKADEMIE DER TECHNIKWISSENSCHAFTEN

Inhalt

Zusammenfassung

Die Mobilität der Zukunft – nutzerfreundlich, ressourceneffizient und individuell flexibel. Hierbei kommt auch Künstlicher Intelligenz (KI) eine wichtige Rolle zu. Denn mit zunehmendem Einsatz von Fahrassistenzsystemen wie Spurhalte- und Überholassistenten bis hin zum autonomen Fahren unter definierten Voraussetzungen wird KI auf unseren Straßen alltäglich. Und auch in anderen Bereichen der Mobilität, etwa bei der Routenplanung im öffentlichen Nahverkehr oder bei autonomen Systemen für die Schiene, schreitet diese Entwicklung voran.

Entscheidend für die erfolgreiche Einführung von KI-gestützter Mobilität ist jedoch Vertrauen – vor allem seitens der Nutzerinnen und Nutzer, denen die technische Umsetzung meist verborgen bleibt. Um Vertrauen zu gewinnen und aufzubauen, bedarf es allerdings mehr als Sicherheit, Transparenz, Datenschutz, Ethik, gesellschaftliche Akzeptanz sowie Regulierung. Denn Vertrauen basiert im Wesentlichen auf persönlichen Erfahrungen, Emotionen und Erlebnissen: Erst im Erleben und Erfahren des eigenen Handelns lernen wir zu verstehen – worauf sich Vertrauen letztlich gründet. Der Schlüssel für Vertrauen in neue KI-basierte Technologien im Bereich der Mobilität und damit in deren Zukunft liegt somit in der untrennbaren Einheit der eigenen wie gesellschaftlichen Akzeptanz, was die aktive Einbindung der Nutzerinnen und Nutzer unabdingbar voraussetzt.

Expertinnen und Experten der Plattform Lernende Systeme unter Federführung der Arbeitsgruppe Mobilität und intelligente Verkehrssysteme mit Beteiligung von Mitgliedern der Arbeitsgruppe IT-Sicherheit, Privacy, Recht und Ethik der Plattform Lernende Systeme beleuchten im Whitepaper das Thema Vertrauen in KI-basierte Mobilität – dies aus den Perspektiven von Anwendenden und Gesellschaft, Unternehmern und Anbietern sowie Institutionen und Politik. Sie untersuchen, wie Vertrauen entsteht, welche Rolle Ethik dabei spielt und wie Erklärbarkeit von KI im Einsatz Vertrauen fördern kann. Das Whitepaper zeigt, dass Vertrauen als Schlüssel zur Akzeptanz neuer KI-Technologien in der Mobilität durch nutzbringende Anwendungen und gesellschaftliche Partizipation aufgebaut werden kann. Der gesamtgesellschaftliche Diskurs sollte dabei an die Informationsbedürfnisse der jeweiligen Zielgruppen angepasst werden, sodass jede durch spezifische Maßnahmen ihren Beitrag erbringen kann, dies als „stimmiger Dreiklang“, wie in den Gestaltungsoptionen dargestellt.

1. Einleitung

Die Mobilität hat sich zu einem zentralen Element der persönlichen Lebensgestaltung entwickelt, und ihre individuelle Ausgestaltung durch Auswahl bevorzugter Verkehrsmodi wird als elementarer Bestandteil der persönlichen Freiheit verstanden. Eine wesentliche Herausforderung ist es, das Mobilitätssystem, wie viele andere Sektoren auch, von einem ressourcenintensiven Sektor hin zu einem ressourceneffizienten Feld wandeln zu müssen, wobei als Ressourcen einerseits Energie und Energieträger zu betrachten sind, andererseits aber auch der öffentliche Raum. Ein wichtiger Baustein dieser Entwicklung ist die Kombination verschiedener Verkehrsträger aus reduziertem Individualverkehr und nutzerfreundlichem öffentlichen Verkehr, dies auch mit neuartigen Fahrzeugkonzepten. Insgesamt bedarf es der Integration verschiedener Aspekte eines intelligenten Zusammenspiels, um eine nutzerfreundliche und nachhaltige Mobilität zu erreichen, die individuelle Flexibilität ermöglicht und gleichzeitig ressourceneffizient ist.

Parallel dazu verbessert die stetige technologische Entwicklung kontinuierlich den Komfort, die Verfügbarkeit sowie die Sicherheit der Mobilitätssysteme: Digitale Navigationslösungen auf Basis von Satellitennavigation für den Individualverkehr, besser abgestimmte Fahrpläne im öffentlichen Straßen- und Schienenverkehr durch den Einsatz digitaler Werkzeuge oder die intelligente, prädiktive Wartung von Zügen auf Grundlage von Sensordaten in den Fahrzeugen sind hierfür heute bereits alltägliche Beispiele. Die Zukunftsvision des autonomen beziehungsweise hochautomatisierten Fahrens rückt durch den verstärkten Einsatz von Fahrassistenzsystemen für routinemäßige Fahrabläufe wie Spurhalte- und Überholassistenten, von Abbiegeassistenten oder automatischen Notbremssystemen und Abstandsregeltempomaten sowie von kamerabasierten Speed Limitern immer näher, wobei sich mit zunehmendem Autonomie- und Automatisierungsgrad auch der Anspruch an Sicherheit, Energieeffizienz und an den Fahrkomfort für den Menschen erhöht. Auf der Schiene und zu Wasser gibt es vergleichbare Entwicklungen mit dem erklärten Ziel, einen hochautomatisierten bis fahrerlosen Betrieb zu ermöglichen, um den öffentlichen Personen- und Güterverkehr flexibler, zuverlässiger, effizienter, inklusiver und nachhaltiger zu gestalten. Letztlich ist der Anspruch, eine vollautomatisch erstellte multimodale Reise- oder Transportkette zu realisieren, die zudem eine echte Alternative zum nach wie vor häufig alleinig präferierten motorisierten Individualverkehr auf der Straße darstellt.

Die Digitalisierung mit der Nutzung von Daten und Mehrwertdiensten ist elementarer Bestandteil dieser Entwicklung. Durch die wachsende Komplexität wird es allerdings zunehmend schwieriger, diese Lösungen allein mit klassisch manuell programmierter Software zu realisieren. Bereits heute werden häufig Methoden der „Künstlichen Intelligenz" (KI) als wichtiges Element der Systeme betrachtet. KI ist dabei nicht nur eine weitere Softwaredisziplin, sondern erweitert die technischen Möglichkeiten, um zu mehr Sicherheit und Lebensqualität beizutragen sowie gleichzeitig den Personen- und Güterverkehr nachhaltig und wirtschaftlich zu gestalten.

Im Gegensatz zu bisherigen klassisch manuell geschriebenen Programmen erfolgt die KI-basierte Entwicklung durch die Bereitstellung von Daten und durch Lernverfahren erzeugte Softwarebausteine auf Grundlage dieser Daten. So entsteht die Situation, dass das Programmverhalten in weiten Teilen maschinell definiert wird und nur indirekt durch den Menschen, über Bereitstellung von Daten und Postulierung von Optimierungskriterien. Dementsprechend gibt es tendenziell keine explizit direkte Verbindung mehr zwischen einer durch Menschenhand entstehenden technischen Spezifikation und einer ebenfalls durch Menschenhand geschriebenen Softwarelösung. Die Anpassung an diese schnellen Entwicklungen stellt eine Herausforderung für das Mobilitätssystem sowie seine Nutzerinnen und Nutzer dar. In diesem ohnehin dynamischen Umfeld sollte der Einsatz von Methoden der Künstlichen Intelligenz so gestaltet sein, dass Vertrauen in diese aufgebaut werden kann.

Vertrauen als zentrales Element

Für die Akzeptanz neuer KI-basierter Technologien in der Mobilität ist neben dem notwendigen Nutzen der Aufbau von Vertrauen der wesentliche Baustein. Dieses Vertrauen muss einerseits von den einzelnen Nutzerinnen und Nutzern des Mobilitätssystems, andererseits auch seitens der Gesellschaft als Ganzes entwickelt werden, um Akzeptanz für neue Mobilitätsformen zu schaffen und das Potenzial von KI-Methoden in dieser Domäne zu erschließen. Fehlendes Vertrauen birgt dagegen die Gefahr, dass Lösungen nicht angenommen und Entwicklungen verlangsamt, eingeschränkt oder verhindert werden.

KURZINFO

Die allgemeine Definition des Begriffs „Vertrauen" lautet: „festes Überzeugtsein von der Verlässlichkeit, Zuverlässigkeit einer Person, Sache" (Duden online: Vertrauen). Wendet man diese Definition auf technische Systeme an, so stellt sich unmittelbar die Frage, was diese Zuverlässigkeit und Verlässlichkeit zum Ausdruck bringen kann, um eine **vertrauenswürdige** Technologie zu schaffen.

In den Ethik-Leitlinien für eine vertrauenswürdige KI der Expertengruppe für Künstliche Intelligenz der Europäischen Kommission ist formuliert, dass sich eine vertrauenswürdige KI durch drei Komponenten auszeichnet, welche während des gesamten Lebenszyklus des Systems sichergestellt sein sollten:

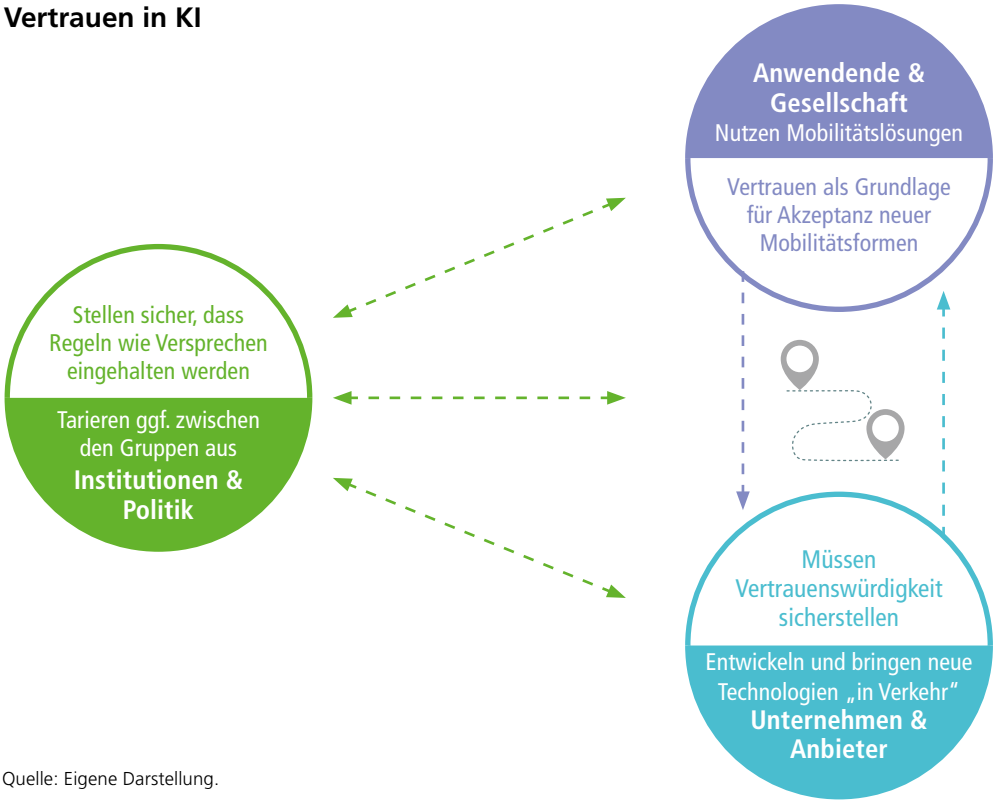
- (1) rechtmäßig: Beachtung aller geltenden Gesetze und Vorschriften
- (2) ethisch: Beachtung der ethischen Grundsätze und Werte
- (3) robust: sowohl aus technischer Sicht als auch unter Berücksichtigung des sozialen Umfelds (European Commission, 2019)

Auf Basis dieser Komponenten werden in den Ethik-Leitlinien sieben Anforderungen zur Verwirklichung einer vertrauenswürdigen KI entwickelt, welche „zur Gestaltung einer kohärenten, vertrauenswürdigen und menschenzentrierten KI" beitragen. Diese Anforderungen wurden auch im Artificial Intelligence Act (AI Act) aufgegriffen und erläutert (European Parliament, 2024; Plattform Lernende Systeme, 2024).

In dieser Publikation soll der Blick auf die verschiedenen Zielgruppen gelegt werden, für die vertrauenswürdige KI für Anwendungen in der Mobilität notwendig ist. Zuvörderst ist es wichtig anzuerkennen, dass „Vertrauen" stark subjektiv geprägt ist und nur indirekt mit der notwendigen Sicherheit technischer Anwendungen in Bezug steht. Weiterhin, dass sich Vertrauen in verschiedenen gesellschaftlichen Gruppen unterschiedlich ausbildet, dass vertrauensbildende Maßnahmen unterschiedlich wirken und dass absolute, allgemeingültige Strategien schwer zu formulieren sind. Während beispielsweise Nutzerinnen und Nutzer einer KI-basierten Mobilitätslösung vertrauen – solange diese zuverlässig zu funktionieren scheint und einen sicheren Eindruck macht – und eher durch die Darstellung aller verfügbaren Informationen verunsichert wären, ist bei einer technischen Prüfung oder einem Audit jedoch der alleinige Nachweis der Abwesenheit von Fehlern kein Beweis für die Vertrauenswürdigkeit eines Systems. Selbst wenn – hypothetisch betrachtet – die Vertrauenswürdigkeit eines Systems mathematisch beweisbar wäre, würde die so nachgewiesene Vertrauenswürdigkeit nicht automatisch zu Vertrauen in allen gesellschaftlichen Gruppen führen.

Um diesen Sachverhalt zu berücksichtigen, werden folgende drei Zielgruppen beleuchtet: Zunächst die Anwendenden und die Gesellschaft, welche stets im Mittelpunkt der Diskussion stehen sollten, da sie die diskutierten Mobilitätslösungen in ihrem Alltag aktiv einsetzen. Daneben die Unternehmer und Anbieter, welche die Technologien überhaupt erst ermöglichen, sowie die Institutionen und die Politik, welche die in der Mobilität unvermeidbaren Zielkonflikte möglichst optimal ausbalancieren müssen.

Abbildung 1: **Unterschiedliche Zielgruppen bei der Diskussion um Vertrauen in KI**



Quelle: Eigene Darstellung.

Drei Leitfragen zu Vertrauen in KI-basierte Mobilität

Das Thema „Vertrauen in KI-basierte Mobilität" betrachtet diese Publikation anhand folgender Fragestellungen:

Leitfrage 1: Wie entsteht Vertrauen?

Die erste Fragestellung widmet sich der Entstehung von Vertrauen. Für die Zielgruppe der Nutzerinnen und Nutzer ist zu erwarten, dass Vertrauen durch wiederholtes positives Erleben entsteht, sofern die jeweilige Mobilitätslösung dabei als sicher empfunden wird und das Verhalten der Mobilitätslösung zumindest in Grundzügen nachvollziehbar ist (zum Beispiel „Geschwindigkeit wurde aufgrund unübersichtlicher Verkehrslage reduziert"). Allerdings ist die Erwartungshaltung an eine Anwendung subjektiv und von Person zu Person individuell unterschiedlich. Daher ist es eine Herausforderung, Verhalten allgemein antizipierbar und gleichzeitig individuell und nachvollziehbar zu gestalten. Hier ist Erfahrung nötig: Reallabore, Pilotprojekte etc. helfen, Vertrauen bei den Nutzenden aufzubauen. Unterschiedliche Domänen sind hier unterschiedlich herausfordernd. Beispielsweise haben schienenbasierte Mobilitätslösungen den Vorteil, dass sie weniger stark in den gemeinsamen Mobilitätsraum integriert sind als etwa der motorisierte Individualverkehr.

Leitfrage 2: Welche Rolle spielt Ethik für Vertrauen?

Ethische Fragestellungen bezüglich der Entwicklung intelligenter Systeme betreffen sowohl die entwickelnden Hersteller und Anbieter als auch die regulierenden Institutionen sowie die Gesellschaft als Ganzes. Vielfach fehlt aber in der Gesellschaft Vertrauen in diese KI-Technologie mit der Konsequenz, dass auf Seiten der Politik Positionen an Popularität gewinnen, welche die Entwicklung einschränken wollen. Gleichzeitig wird im Kontext der Ethik-Leitlinien der EU eine vertrauenswürdige KI implizit durch die drei zentralen Komponenten Rechtmäßigkeit, Ethik und Robustheit qualifiziert (siehe Infokasten: Vertrauen). Ziel der Diskussion um ethi-

sche Aspekte von KI in der Mobilität muss es sein, dass die Chancen und Risiken unter den Betroffenen fair verteilt sind. Gerade bei dieser Fragestellung ist es wichtig, eine konstruktive Diskussion in der gesamten Gesellschaft zu erreichen, um den Einsatz von KI-basierten Mobilitätslösungen mit dem Rückhalt der Bevölkerung zu entwickeln.

Leitfrage 3: Welche Rolle spielt Erklärbarkeit von KI, um Vertrauen zu gewinnen?

Klassisch (das heißt ohne hochperformante KI-Methoden) entwickelte und programmierte KI-Systeme beziehen Vertrauen stark daraus, dass Menschen mit entsprechender Expertise das Wirkprinzip des Systems im Detail erklären können. Im Kontext der neuen, hochautomatisierten (KI-)Anwendungen und der entsprechenden heute erforschten und entwickelten KI-basierten Lösungen sind Erklärungen im technischen Detail oft nicht durchgängig beziehungsweise nur auf Kosten von Leistungsfähigkeit möglich. Es stellt sich daher die Frage, wie man eine Abwägung der Aspekte Erklärbarkeit und Leistungsfähigkeit in Bezug auf Vertrauen bilden kann. Wir werden aufzeigen, dass der Begriff „Erklärbarkeit" eine Reihe von Teilaspekten beinhaltet, die es lohnt, gesondert zu betrachten.

USE CASE

Hochautomatisiertes Fahren als ein anschauliches Beispiel für vertrauenswürdige KI

Für die weiteren Ausführungen ist zu beachten, dass die Publikation verschiedene Bereiche der Mobilität beleuchtet, beispielsweise:

- ein multimodales Mobilitätssystem, das durch KI optimiert wird
- durch KI erstellte „Digitale Zwillinge" eines solchen Mobilitätssystems

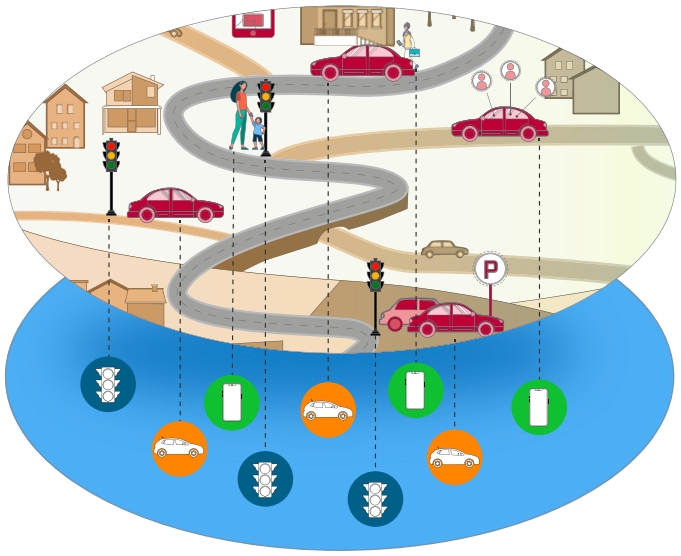


Abbildung: Digitaler Zwilling – Ein digitaler Zwilling ist das digitale Abbild wesentlicher Elemente eines Systems aus der realen Welt.

- ein KI basiertes Wartungs- und Instandhaltungssystem von z. B. Zug- oder Bussystemen
- Funktionen des autonomen oder hochautomatisierten Operierens von Automobilen, Zügen, Schiffen oder im Flugverkehr.

Für eine fokussierte und durchgängige Diskussion wird an verschiedenen Stellen das **Beispiel des hochautomatisierten Fahrens** intensiver herangezogen. Die Argumentationen passen allerdings häufig auch auf andere Felder der Mobilität und sollten auf diese gegebenenfalls projiziert werden können.

2. Wie entsteht Vertrauen?

Damit das Potenzial KI-basierter Mobilitätsanwendungen ausgeschöpft werden kann, bedarf es des Vertrauens in die Anwendungen und in deren Vorteile, insbesondere seitens der Nutzerinnen und Nutzer. In der einfachsten Betrachtung entsteht Vertrauen durch wiederholtes positives Erleben und Erreichen übereinstimmender Ziele. Wenn sich eine KI-basierte Anwendung kontinuierlich so verhält, wie Anwendende dies erwarten, wird dieser Anwendung Vertrauen geschenkt. Wenn autonome Systeme Ziele verfolgen, die mit denen der Nutzenden übereinstimmen, wie beispielsweise erhöhte Sicherheit, kürzere Reisezeiten oder geringerer Energieverbrauch, werden diese positiv erlebt und daher mit höherer Wahrscheinlichkeit akzeptiert (Choi & Ji, 2015).

Gleichzeitig muss sichergestellt werden, dass keine überhöhten oder unrealistischen Erwartungen an die KI-basierten Systeme und ihre Fähigkeiten gestellt werden. Beispielsweise könnte die Erwartung, dass Unfälle unter Beteiligung teilautonom fahrender Fahrzeuge absolut ausgeschlossen seien, zu einem bewusst (Hungund & Pradhan, 2021) oder unbewusst (Lyu et al., 2020) weniger umsichtigen Verhalten im Straßenverkehr führen (DEKRA, 2023). Dies würde die Risiken erhöhen und die positiven Effekte einer unter Beibehaltung der ursprünglichen Umsichtigkeit durchaus erhöhten Sicherheit aufheben.

Wenn die Interaktion zwischen Anwendenden und den KI-Systemen nicht gut abgestimmt ist, kann ein Eingriff eine Kompensationsreaktion provozieren und die Vorteile dieses Systems aufheben: Beispielsweise kann ein autonomer Spurhalteassistent die Fahrerin oder den Fahrer überraschen und durch eine manuelle Eingriffsreaktion der Fahrerin oder des Fahrers damit das Unfallrisiko erhöhen. Entspricht der Eingriff des Assistenzsystems nicht der Erwartungshaltung, ist dieser beispielsweise zu früh, zu spät oder nach Einschätzung der Anwendenden unnötig, geht das Vertrauen in die Assistenzsysteme verloren. Ein anderes Beispiel ist eine durch KI-Methoden prognostizierte hohe Auslastung eines öffentlichen Verkehrsmittels, welches Reisende zum Wechsel auf alternative Verkehrsträger bewegt, obwohl es sich lediglich um eine Prognose handelt.

Gleichzeitig ist damit zu rechnen, dass das Abwägen der Erwartungshaltung bezüglich derartiger Situationen stark abhängig von den Nutzenden ist und es eine Herausforderung an sich darstellt. Beispielsweise kann Person A einen defensiven und ruhigen Fahrstil bevorzugen und einen sportlichen Fahrstil als unsicher erleben, während die Erwartungshaltung von Person B genau entgegengesetzt ist. Auch im Falle einer KI-erzeugten Empfehlung, wegen schlechten Wetters das Fahrrad zu meiden, wäre die Wahrnehmung der Angemessenheit dieser Empfehlung stark individuell. Dies bedeutet, dass das Verhalten eines KI-basierten Systems nicht für alle Nutzenden gleichgeschaltet eingestellt werden kann. Vielmehr kann eine gewisse Individualisierung derartiger Einstellungen sinnvoll oder gar notwendig sein. Dies kann dann entweder manuell erfolgen oder seinerseits automatisiert durch den Einsatz adaptiver Systeme, gegebenenfalls erneut unter Verwendung von KI-Methoden.

Die Thematik des Vertrauens in KI-basierte Mobilität hat zusätzlich zu potenziell individuell bedingten Effekten eine gesellschaftlich-kulturelle Komponente. Während manche Gruppen beziehungsweise Kulturen sich Neuem eher skeptisch nähern und innovative Lösungen nur zögerlich annehmen, sind andere bereit, sich schnell auf Neues einzulassen und sich erst im Nachgang, basierend auf gemachten Erfahrungen, ein Urteil über das Neue zu bilden. Um Vertrauen, Nutzung und Akzeptanz von KI-basierten Systemen zu fördern, sind daher Strategien zu entwickeln, die helfen, ein angemessenes Maß an Vertrauen zu erreichen und eine ausgewogene Einstellung zum zweckmäßigen und sicheren Einsatz solcher Systeme zu fördern. Dies gilt für Neues generell, für KI-basierte Systeme im Speziellen und entsprechend auch für KI-basierte Mobilitätsanwendungen. Exemplarisch seinen nachfolgend einige Beispiele genannt.

Schrittweise Einführung der Technologie in weniger kritischen Anwendungsszenarien

Das Vertrauen in den Einsatz und in die Nutzung von Innovationen ist stark von ihrer reibungslosen Einführung abhängig. Nach dem Technology Acceptance Model kann eine neue Technologie die häufig vorhandene anfängliche Skepsis durch eine Kombination aus wahrgenommener Nützlichkeit und wahrgenommener Benutzerfreundlichkeit überwinden (Davis,1985). Eine Strategie, in dem teils hochkomplexen Feld der KI-basierten Mobilität Vertrauen aufzubauen, kann beispielsweise darin bestehen, Anwendungsfälle zu identifizieren und umzusetzen, die zu Beginn bereits einen sicheren Nutzen erlauben, auch wenn dieser Nutzen nur einen Teilbereich eines größeren Zielnutzens erfüllt. So sind heute bereits autonome Fahrfunktionen in eingeschränkten Situationen zugelassen, beispielsweise bis 60 km/h auf Autobahnen oder auf größeren Parkplätzen oder auf Flughäfen in Verbindung mit Fahrgasttransportsystemen. Gelingt es, Anwendungen dieser Art reibungslos und in größerer Anzahl zu realisieren, kann die Weiterentwicklung der Technologie einerseits von den Erfahrungen mit solchen Anwendungen profitieren, anderseits wird mit zunehmender Anzahl solcher Anwendungsbeispiele das Vertrauen in die Technologie kontinuierlich wachsen.

Sicherstellung von Kontrolle durch die Nutzenden

Wenn Nutzende eine neuartige autonome Technologie, wie zum Beispiel KI-gestütztes autonomes Fahren oder KI-basierte Navigation per Smartphone, kennenlernen, kann unter Umständen Misstrauen aufgrund eines Kontrollverlusts über das System entstehen. In solchen Fällen, wie beispielsweise bei einem Eingreifen der KI-basierten automatisierten Fahrfunktionen wie dem Spurhalteassistenten, ist es ratsam, die Möglichkeit der Korrektur durch die Anwendenden selbst sicherzustellen. Zusätzlich ist es sinnvoll, dass diese Korrekturen vom System wahrgenommen und als Präferenz gelernt werden, sofern diese Verhaltensweisen die Sicherheit nicht gefährden und alle geltenden Gesetze und Vorschriften beachten und diese gelernten Verhaltensweisen, generell die Bedingungen für das Verhalten des Systems für die Nutzenden, transparent bleiben. Auch bei der Übermittlung von gesammelten Daten aus dem Fahrzeug oder dem Smartphone muss den Anwendenden die Möglichkeit gegeben werden, die Übermittlung, wie in der Datenschutzgrundverordnung (DSGVO) vorgeschrieben, zu kontrollieren und ihr gegebenenfalls zu widersprechen. Wenn der Grund zur Datenübermittlung und der Nutzen transparent und nachvollziehbar gemacht werden, werden die Nutzenden eher zustimmen, weil „sichtbar" gemacht wird, wie das Sammeln und die Übermittlung der Daten mit ihren Interessen (zum Beispiel Verbesserung des Fahrerlebnisses, der Sicherheit oder der Effizienz) übereinstimmt (Farke et al., 2021)

Die Basis für einen vertrauensvollen Umgang und die wiederholte Nutzererfahrung mit positivem Ergebnis sind nicht primär Erklärung und Wissensvermittlung, sondern die aktive Interaktion zwischen Menschen, Technik und den KI-basierten Funktionen, wobei Technik und KI immer im Sinne des Menschen funktionieren müssen und die KI-basierten Funktionen positiv erlebbar und zuverlässig bleiben sollten. Fehlerhafte oder anpassungsunfähige Systeme bauen dagegen Vertrauen eher ab (Salem et al., 2015).

Die Federal Aviation Administration (FAA) der USA schätzt, dass aktuell in 90 Prozent der Flugzeit Automatisierung genutzt wird, und befürchtet eine Abhängigkeit von der Automatisierung. Es sei unklar, „[…] ob die Piloten ausreichend geschult und erfahren sind, um ihre manuellen Flugfähigkeiten zu erhalten" (U.S. Department of Transportation, Office of the Secretary of Transportation, 2016). Es ist zu erkennen, dass die Nutzung von KI-gestützten autonomen Systemen oder Assistenzsystemen die eigenen menschlichen Fähigkeiten reduziert, was wiederum die Nutzung dieser Systeme erhöhen kann und somit schließlich mehr Vertrauen erzeugt. So zeigte die „Wizard-of-Oz"-Studie (Weiss et al., 2009; Dahlbäck et al., 1993) bei PSA (Schlag, 2016), dass

Passagiere, die in einem vermeintlich voll automatisierten Auto saßen, durch Erfahrungen mit den Systemen Vertrauen und Akzeptanz aufbauen konnten. Bei dieser Entwicklung muss beachtet werden, dass ein Verlust an menschlichen Fähigkeiten zu einem höheren Risiko führen kann, wenn die Technologie diesen Verlust nicht kompensieren kann, was wiederum das Vertrauen in die Technologie schädigen kann. Daher sollte die Einführung autonomer Systeme schrittweise erfolgen, um die Sicherheit jederzeit gewährleisten zu können, und die Fähigkeiten des Systems sollten nicht übertrieben dargestellt werden, um blindes Vertrauen zu vermeiden.

Aufgeklärter gesellschaftlicher Umgang

Der Aspekt des Verfolgens gemeinsamer Ziele zur Schaffung von Vertrauen kann ebenso für die Verbesserung der gesamtgesellschaftlichen Akzeptanz von KI-basierten Mobilitätstechnologien wie dem autonomen Fahren genutzt werden. Hier ist es besonders wichtig, ein realistisches Bild der Sicherheit solcher Systeme in der Gesellschaft entstehen zu lassen, da sowohl marketinggetriebene übertriebene Darstellungen des Reifegrads wie auch medial beeinflusste negative Darstellungen die Einführung solcher Systeme beeinträchtigen können (Jelinski et al., 2021). Wenn gezeigt und kommuniziert werden kann, dass KI-basierte Mobilitätstechnologien wie autonome Fahrzeuge oder KI-basierte Kartennavigation einerseits individuelle Ziele wie angenehmeres, schnelleres, sparsameres oder sichereres Reisen und andererseits gesamtgesellschaftliche Ziele wie höhere Sicherheit im Straßenverkehr und eine verbesserte Luftqualität sowie verbesserte Parkraumnutzung in Städten ermöglichen, können große Gruppen an Nutzenden und letztendlich somit die Gesellschaft diese Technologie positiv erleben, was Akzeptanz und Vertrauen schafft. Wenn diese Aspekte jedoch vernachlässigt werden, kann sich dies dagegen in Vertrauensverlust umkehren, was die Einführung der Technologie negativ beeinflussen würde. Beispielsweise kann beobachtet werden, dass Berichte zu Verkehrsunfällen, bei denen KI-Systeme eine Rolle spielen, medial intensiv aufgegriffen werden und bei vielen Menschen zu negativen Konnotationen der „neuen" Technologie führen. Auch ist zu erwarten, dass Unfälle mit Beteiligung von KI-basierten autonomen Systemen ein ungewohntes Profil haben werden, zum Beispiel Verkehrsteilnehmende mit Gegenständen verwechselt werden. Dies kann den Eindruck entstehen lassen, dass die Systeme nicht ausgereift sind. Hier ist es wichtig, sowohl negative als auch positive Entwicklungen sorgfältig aufzuarbeiten und transparent zu kommunizieren, um die Nutzenden mit der neuen Technologie kontinuierlich vertraut zu machen, den erwarteten Mehrwert der neuen Technologie für die Gesellschaft kritisch zu prüfen und gegebenenfalls zu belegen.

Die Sicherstellung einer unbedenklichen Nutzung einer Technologie wie dem autonomen Fahren wird von gesellschaftlichen Bedürfnissen beeinflusst. Die Technologie kann nur schwer zur Einsatzreife gelangen, wenn sie vor der Inbetriebsetzung nur unter realitätsfernen Bedingungen entwickelt und getestet wird. Stattdessen ist eine stufenweise Eröffnung der Technologie in der Praxis wertvoll, um gesellschaftliche Bedürfnisse mit technischen Anforderungen abzugleichen (OECD/ITF, 2015) und etwaige Zielkonflikte zu verstehen, transparent zu machen und offen zu diskutieren und zu behandeln. Statistische Wahrscheinlichkeiten allein sind nicht ausreichend, um KI-basierte Systeme anzulernen und zu zertifizieren. Erst im Zusammenspiel mit den individuellen Sicherheitsinteressen und Bedürfnissen der Fahrerinnen und Fahrer erlangen sie Vertrauen und Akzeptanz.

3. Welche Rolle spielt Ethik für Vertrauen?

Gesellschaft wie auch Anwendende erwarten von autonomen technischen Systemen, dass sie sich entsprechend den ethischen Grundsätzen der jeweiligen Gesellschaft verhalten. Die Entwicklung von Systemen, die dieser Anforderung Rechnung tragen, wird in dem noch recht jungen Forschungsfeld der Maschinenethik betrachtet. An dieser Schnittstelle von Informatik und Philosophie werden Computersysteme entwickelt, die sich auf der Grundlage vorliegender Daten und Algorithmen so verhalten, dass von „moralischem Verhalten" gesprochen werden kann. Kommen dabei Verfahren der Künstlichen Intelligenz zum Einsatz, spricht man auch von „Artificial Morality" (AM). Auf Nutzerinnen und Nutzer kann das Verhalten dieser Systeme so wirken, als ob „moralische Entscheidungen" getroffen würden. In Wirklichkeit können diese technischen Systeme aber nur Berechnungen durchführen – die Grundlage der Entscheidung basiert darauf, dass die Systeme von den Unternehmen nach ethischen Grundsätzen gestaltet wurden.

Es ist zu erwarten, dass Lernende Systeme in der Mobilität in Situationen sein werden, wo ethisches Verhalten erwartet wird. Beispielsweise sollen autonome Fahrzeuge bei Entscheidungen menschliches Leben höher priorisieren als Sach- und Tierschäden. Häufig werden hier moralische Dilemmata diskutiert, wie das schon vor knapp 75 Jahren formulierte Gedankenexperiment des Trolley-Problems, in dem ein Weichensteller einen Zug umleiten kann, der auf eine Menschenmenge zurast (Foot, 1978).

Die Betrachtung des Trolley-Problems ist allerdings für echte Anwendungsfälle wenig zielführend, da es eine hochgradig konstruierte Situation darstellt, die in der Realität mit verschwindend geringer Wahrscheinlichkeit auftreten wird. Ähnlich wie bei menschlichen Fahrerinnen und Fahrern, die auf eine Situation nicht rational, sondern nur reflexartig reagieren können, ist es die beste Strategie, durch andere Maßnahmen, wie reduzierte Geschwindigkeit, ein solches Dilemma gar nicht erst entstehen zu lassen. In diesem Kontext sind die ethischen Grundsätze im Straßenverkehr bereits durch die Straßenverkehrsordnung definiert (Bundesministerium der Justiz, StVO §1). Wenn ein Fahrzeug dieses Regelwerk befolgt, kann von einem ethischen Verhalten gesprochen werden. Sollten sich autonome Fahrzeuge strenger an die Straßenverkehrsordnung halten als menschliche Fahrerinnen und Fahrer, wären diese damit auch potenziell „ethischer" unterwegs.

Dagegen sind andere Dilemmata sehr wohl für KI relevant und könnten schon bald in deren Entscheidungsbereich fallen. Beispielsweise muss im öffentlichen Personenverkehr täglich abgewogen werden, ob eine pünktliche Abfahrt hinausgezögert werden kann, um unverschuldet verspäteten Reisenden eines gerade ankommenden Fahrzeugs noch den Zustieg und damit die Fortsetzung ihrer Reise zu ermöglichen. Dieses Dilemma kann dem Ethikbereich der Verteilungsgerechtigkeit zugerechnet werden und kann im interregionalen Schienenverkehr ebenso auftreten wie an einer Bushaltestelle, wo einzelne Personen auf den abfahrenden Bus zueilen. Bisher wird diese Entscheidung von einzelnen Personen auf Grundlage vieler Parameter (zum Beispiel Lage der Haltestelle, Uhrzeit, Anzahl und gegebenenfalls eingeschränkte Mobilität der Reisenden, Toleranzen im Fahrplan) intuitiv getroffen. Autonome Systeme werden in Zukunft angemessen reagieren müssen. Dieses Verhalten wird daher großen Einfluss auf das Vertrauen der Reisenden in diese Systeme haben. Derartige Fragestellungen befinden sich in einem Spannungsfeld zwischen Geistes-, Sozial-, Natur- und Ingenieurwissenschaften, weshalb Lösungsansätze interdisziplinär erarbeitet werden müssen.

Wie bereits beschrieben, liegt die Verantwortung für das Verhalten eines autonomen Systems bei allen Personen, die an der Entwicklung beteiligt sind. Auch bei neuartigen selbstlernenden Systemen, welche die Fähigkeit haben, sich auf Grundlage von Trainings- beziehungsweise Lerndaten autonom weiterzuentwickeln und sich in neuen Situationen autonom zurechtzufinden, muss aus ethischer Sicht die Verantwortung weiterhin bei den für die Entwicklung verantwortlichen Personen und nicht bei den Halterinnen und Haltern oder Fahrerinnen und Fahrern liegen.

Die Ethik-Leitlinien für vertrauenswürdige KI

Die EU hat zu diesem Thema eine hochrangige Gruppe von Expertinnen und Experten für Künstliche Intelligenz ins Leben gerufen, die dazu „Ethik-Leitlinien für eine vertrauenswürdige KI" entworfen hat. Diese hat formuliert, dass KI (1) rechtmäßig, (2) ethisch, (3) robust sein soll (Hochrangige Expertengruppe für Künstliche Intelligenz, 2018). Dies waren auch die Leitgedanken bei der Diskussion und Entstehung des AI Acts der EU.

„Rechtmäßig" bedeutet gemäß den Ethik-Leitlinien, dass sich die KI auf Basis des geltenden Rechts verhält. Dabei sind sowohl das EU-Primärrecht und das EU-Sekundärrecht als auch die Verordnungen der einzelnen Mitgliedsstaaten zu beachten. Das Verhalten der KI wird hier aus zwei Dimensionen betrachtet, zum einen, was klar verboten ist, zum anderen, was ausdrücklich erwünscht ist.

„Ethisch" bedeutet, dass die KI sich an grundlegende ethische Prinzipien und Werte hält. Die Expertinnen und Experten haben dazu vier ethische Grundprinzipien aufgestellt, die unbedingt zu beachten sind: „Achtung der menschlichen Autonomie", „Schadensverhütung", „Fairness" und „Erklärbarkeit". Auf diese Grundprinzipien kann immer zurückgegriffen werden, wenn beispielsweise die rechtliche Grundlage nicht zur Lösung eines Problems geeignet ist. Dies kann der Fall sein, wenn die technologische Entwicklung schneller fortschreitet als die Gesetzgebung.

„Robust" steht laut der Expertinnen und Experten sowohl im technischen als auch im sozialen Zusammenhang. Technisch gesehen soll die KI in unerwarteten Situationen robust sein, also kein unerwünschtes Verhalten zeigen, was sonst eine Gefahr für das soziale Umfeld, in dem sie operiert, darstellen könnte.

Alle drei Komponenten sind als wesentliche Elemente einer vertrauenswürdigen KI anzusehen. Um dies in der Praxis auch erfolgreich umsetzen zu können, wird allerdings die Zusammenarbeit innerhalb der Gesellschaft benötigt.

KURZINFO

Diskurs von Ethik und Technik

Die Diskussion über eine vertrauenswürdige KI findet oft isoliert von den aktuellen technischen Möglichkeiten statt. Da aber alle ethischen Anforderungen an eine KI letztendlich auch von Entwickelnden und Unternehmen umgesetzt werden müssen, dürfen diese Perspektiven nicht außer Acht gelassen werden. Häufig offenbart sich ein Konflikt zwischen den theoretisch diskutierten Szenarien und Forderungen und der praktischen Umsetzung, insbesondere im ökonomischen Kontext. Um diese Diskrepanzen zu überwinden und ein breites Vertrauen in KI-Anwendungen zu entwickeln, ist ein aufgeklärter gesamtgesellschaftlicher Dialog auf Augenhöhe über Chancen und Risiken des Einsatzes von KI notwendig.

In der öffentlichen Diskussion ist, besonders in Bezug auf autonomes Fahren, gerade nach Unfällen die Berichterstattung durch einen negativen Gesamtton geprägt (Jelinski et al., 2021). Wie im vorherigen Abschnitt beschrieben, sollte angestrebt werden, diese subjektive Einschätzung mit dem objektiven erwarteten Sicherheitsgewinn durch autonome Fahrzeuge in der Praxis in Bezug zu setzen, um den Nutzerinnen und Nutzern eine souveräne Konsumentenentscheidung über die Verwendung der Technologie zu ermöglichen. Es sollte transparent dargestellt werden, wie die Sicherheitsanforderungen an diese Fahrzeuge den ethischen Herausforderungen Rechnung tragen. Fragen zur Maschinenethik sollten allerdings weiterhin Bestandteil des Diskurses über den Einsatz Lernender Systeme für autonomes Fahren sein, dessen übergeordnetes Ziel die realistische Abwägung verschiedener Faktoren wie Sicherheit, Geschwindigkeit und Ressourcenverbrauch im Verkehrssystem ist. Es sollte versucht werden, die möglichen Vorteile eines KI-Einsatzes nicht durch die Dominanz des Aspekts „Angst" zu verfälschen.

Durch die breite Beteiligung an der Diskussion dieser ethischen und technischen Fragen wird deutlich, dass Konsens und Verantwortung gesamtgesellschaftlich getragen werden sollten. Das wirft unmittelbar die Notwendigkeit auf, bei Entwicklungen von KI-basierten Systemen kulturelle und ethische Unterschiede in verschiedenen Regionen der Welt zu berücksichtigen, wie dies auch bei anderen technischen Systemen notwendig ist. Ein signifikanter Unterschied besteht darin, ob Entscheidungen situativ getroffen werden oder im Voraus vorgegeben sind. Ein weiterer Aspekt ist das potenzielle Auftreten von Lücken in der Verantwortung für Entscheidungen beziehungsweise in der Verantwortung der Entwicklung von einzelnen autonomen Systemen (Fiktives Gerichtsverfahren, Plattform Lernende Systeme, 2022). KI ist dabei nicht per se inhärent ethisch oder unethisch – dies entscheidet sich durch ein Einsatzszenario, in welchem die KI von den Entwickelnden als technische Lösung eingesetzt wird. Daher ist es unumgänglich, für den Einsatz Anforderungen zu definieren beziehungsweise diese gesetzlich vorzuschreiben (DKE & DIN, 2022).

Der Weg, um das Vertrauen in KI-Systeme und deren ethisches Verhalten zu stärken, ist, diese Systeme vertrauenswürdig und zuverlässig zu gestalten. Dies kann gelingen, wenn bei der Entwicklung solcher Systeme nicht allein die Technik im Mittelpunkt steht, sondern deren Integration und Anwendung im jeweiligen Anwendungsbereich, hier konkret im Mobilitätssystem.

4. Welche Rolle spielt Erklärbarkeit von KI, um Vertrauen zu gewinnen?

Ein weiterer wichtiger Aspekt, um die Vertrauenswürdigkeit KI-basierter Mobilitätslösungen zu steigern, ist eine Nachvollziehbarkeit der durch KI getroffenen Entscheidungen. Diese kann dadurch erreicht werden, indem eine Entscheidung durch das System nachvollziehbar erklärt wird, wobei die Herleitung und Erläuterung für die Nutzerin und den Nutzer weniger relevant ist als die Nachvollziehbarkeit der Entscheidung selbst. Zwar sind KI-Systeme, die Erklärungen geben können, nicht automatisch vertrauenswürdig, jedoch werden dadurch Verständlichkeit und Transparenz gesteigert, was zu einer realistischen Einschätzung der Fähigkeiten und Limitierungen von KI-Systemen führen kann und somit zu angemessenem Vertrauen (Wang et al., 2016a; Wang et al., 2016b).

Erklärbarkeit von KI und Vertrauen in KI

Bei klassischen Produkten ist neben Faktoren wie der empfundenen und tatsächlichen Qualität, der Reputation des Herstellers, dem Design, dem bewährten Einsatz im sozialen Umfeld die Erklärung durch einen anderen Menschen, beispielsweise im Einzelhandel, ein wichtiges Element für Vertrauenswürdigkeit. Bei von Menschen klassisch programmierter Software ist die Erklärung der Entwickelnden oder im kommerziellen Bereich ein Interview in Form eines Audits eine wichtige Hilfe für die Vertrauensbildung beim Kunden. Ist die Funktionsweise des Systems durch die Erklärung nachvollziehbar und das Verhalten erwartbar, dann wird das System im Allgemeinen auch als vertrauenswürdig erlebt, weil das Verhalten des Systems nachvollziehbar und vorausschaubar ist. Daher hat sich bei klassischen Systemen das erklärbare Verhalten als Stütze für Vertrauenswürdigkeit etabliert.

Für KI-basierte Produkte bedeutet dies, dass, je mehr Nutzende die Möglichkeit haben, KI zu erleben und je mehr sie dadurch nachvollzogen werden kann, sprich, je besser sie sich erklären wird, desto höher wird das Vertrauen der Nutzenden darin sein. Werden KI-Lösungen jedoch angeboten, um über die dadurch entstandenen Nutzerinteraktionen intrasparente Zweitnutzungen für den Anbieter zu erzielen – zum Beispiel Geschäftsmodelle auf Basis der durch die Anwendung erzeugten Daten –, wird das Vertrauen auch wieder verschwinden beziehungsweise gar nicht erst entstehen können.

Für Anwendungen im Bereich der Mobilität bedeutet dies, eine robuste, erklärbare und erlebbare KI zu verwenden! Und dies zudem stets aus dem gemeinsamen Interesse heraus, der Vision einer möglichst unfallfreien und nachhaltigen Mobilität näher zu kommen. Fahrerloses Fahren wird ohne KI nicht realisierbar sein und braucht daher eine Prüfung bezüglich Verlässlichkeit und Kontrollmechanismen, um zumindest das gleiche Niveau von Vertrauen und Sicherheit zu erreichen, wie es heute von menschlichen fahrenden Personen verlangt wird. Vertrauen in die Reife von fahrerlosen Systemen entsteht unter anderem durch langfristige Tests, aber auch durch eine gute Erklärbarkeit der Funktion, der Reaktionen und Fahrentscheidungen in einzelnen Situationen.

Zwar wird das Verhalten von menschlichen Autofahrerinnen und Autofahrern, das oft intuitiv und somit nicht immer rational erklärbar ist, toleriert, und sofern das System „Mobilität" in Kombination mit dem menschlichen Verhalten (teil-)autonom funktioniert, dann auch reichlich genutzt. Allerdings bietet der Einsatz von erklärbaren autonomen Systemen ein großes Verbesserungspotenzial des Status quo, einerseits als Daten-

grundlage für a-posteriori-Analysen, zum Beispiel zur Erklärung eines Unfalls, wie dies in der Luftfahrt üblich ist, aber auch zur besseren Erlebbarkeit der Nutzerinnen und Nutzer während der Reise, um das Verhalten eines Fahrzeugs nachvollziehbar gestalten zu können und damit das Vertrauen in Assistenzsysteme oder autonome Funktionen zu stärken.

Die Bedeutung der Erklärbarkeit für die Vertrauensbildung hat unter anderem auch die Europäische Kommission in ihren „Ethik-Leitlinien für eine vertrauenswürdige KI" herausgestellt (European Commission, 2019). Erklärbarkeit wird unter der Überschrift der Transparenz als eine von sieben Kernanforderungen für die Verwirklichung einer vertrauenswürdigen KI aufgeführt. Hier wird eine kontinuierliche Bewertung und Berücksichtigung aller Kernanforderungen, und damit auch der Erklärbarkeit, während des gesamten Lebenszyklus des KI-Systems gefordert.

Auch andere Regionen und Länder haben in ihren Regulierungen Erklärbarkeit aufgenommen, wie zum Beispiel die USA im Algorithmic Accountability Act 2019 (Clarke, 2019). Darin werden Unternehmen verpflichtet, eine Bewertung der Risiken vorzunehmen, die das automatisierte Entscheidungssystem für die Privatsphäre oder die Sicherheit darstellt, sowie der Risiken, die zu ungenauen, unfairen, voreingenommenen oder diskriminierenden Entscheidungen beitragen, die sich auf Verbraucherinnen und Verbraucher auswirken.

Inwieweit kann KI sich selbst erklären? In jüngster Zeit haben vermehrt statistische und datengetriebene Ansätze im maschinellen Lernen (ML) eine hohe Effektivität bei der Lösung komplexer Aufgabenstellungen, insbesondere bei hochautomatisierten Systemen, gezeigt (Krizhevsky et al., 2012). Bei aller Leistungsfähigkeit haben diese Ansätze den Nachteil, dass das Eingabe-Ausgabe-Verhalten für einen Menschen nicht mehr direkt und einfach nachvollziehbar oder erklärbar ist und mitunter überraschend auftreten kann. Diese Systeme mit hochdimensionalen Parameterstrukturen werden als undurchsichtige „black-boxes" vom Menschen wahrgenommen. Selbst für Entwicklerinnen und Entwickler ist das Verhalten in der Regel nicht im Detail nachvollziehbar oder erklärbar. Wie oben erläutert, kann die fehlende Transparenz vertrauensmindernd und damit problematisch sein, insbesondere in sicherheitskritischen Anwendungen.

Mittlerweile existieren Methoden, um aus einem KI-basierten System Erklärungen zu extrahieren. Diese Methoden werden als „Erklärbare KI" (englisch „Explainable AI", kurz „XAI") bezeichnet. Allerdings geht eine bessere Erklärbarkeit in der Regel zu Lasten der Leistungsfähigkeit dieser Systeme (Molnar, 2020). Mit dem Fortschritt der ML-basierten Lösungsansätze wird durch die Technologie deutlich an Performanz gewonnen. Ohne zusätzliche Maßnahmen verliert man aber zunächst an Möglichkeiten der Erklärbarkeit. Ihre Realisierung ist möglich, stellt aber für die Systeme einen erheblichen, ressourcenaufwändigen Mehraufwand dar, insbesondere wenn eine hohe Qualität auch in schwer vorhersagbaren Ausnahmesituationen im Straßenverkehr gefordert ist oder sie zudem auch für Laiinnen und Laien verständlich sein soll. Zwischen Leistungsfähigkeit und Erklärbarkeit besteht in der Regel ein Zielkonflikt, der beim Einsatz von XAI individuell abgewogen werden muss. Daher sollte eine Forderung nach Erklärbarkeit immer mit dem dafür nötigen Aufwand bei Entwicklung und Betrieb in Relation gesetzt werden.

Für Nutzerinnen und Nutzer reicht in den meisten Fällen eine grobe Erklärung des Verhaltens aus. Insbesondere während der Nutzung der jeweiligen Verkehrsmittel wären ausführliche Erklärungen der jeweiligen Entscheidungen zu überfordernd. Für Unternehmen und Entwicklerinnen und Entwickler sowie für Institutionen hingegen kann Erklärbarkeit sehr wertvoll sein, um das Verhalten des Systems zu analysieren und nachvollziehen zu können. Seit dem Aufkommen der ersten neuronalen Netze haben Forscherinnen und Forscher, Ingenieurinnen und Ingenieure und Fachleute das Bedürfnis, ihre komplexen, nichtlinearen Modelle zu verstehen. Während in den frühen Tagen der KI-Forschung darauf abgezielt wurde, Verbindungen zwischen

KI-Modellen und der menschlichen Neurodynamik zu finden, konzentrierten sich spätere Arbeiten mehr auf das Verständnis der erlernten Repräsentationen und des Verhaltens des Systems, zum Beispiel durch die Extraktion von Regeln aus neuronalen Netzen oder deren Visualisierung mithilfe sogenannter „Saliency Maps", welche die Eingabefragmente hervorheben, die für die Entscheidung des neuronalen Netzes von hoher Relevanz waren. Mit dem Aufkommen von Deep Learning wurde der Wunsch nach KI-Transparenz noch stärker. Der zunehmend breitere Einsatz von KI-Systemen auch für sensible Anwendungen (neben den sicherheitskritischen Anwendungen in der Mobilität auch zum Beispiel im medizinischen Bereich) oder als Werkzeug zur Datenverarbeitung in Wissenschaft und Wirtschaft wird die Nachfrage nach erklärbaren KI-Methoden (sogenanntes XAI) voraussichtlich weiter erhöhen.

Neuronale Netze sind von der Art und Weise inspiriert, wie menschliche Neuronen und Synapsen lernen, indem Verbindungen gebildet und gestärkt werden. Trainingsdaten wie Bilder oder Audio werden in ein neuronales Netzwerk eingespeist, das nach und nach angepasst wird, bis es richtig reagiert. Ein Deep-Learning-Programm kann trainiert werden, um Objekte in Fotos mit hoher Genauigkeit zu erkennen, vorausgesetzt, es sieht viele Trainingsbilder und erhält viel Rechenleistung. Deep Learning ist gut darin, Muster in Unmengen von Daten zu finden, kann aber nicht erklären, wie sie miteinander verbunden sind. Für Menschen ist es möglicherweise auch nicht sofort ersichtlich, was die Bilder gemeinsam haben, sodass überraschende Ergebnisse zutage treten können. Für wenig kritische Anwendungsbereiche, wie zum Beispiel Unterhaltung über Musik- oder Filmempfehlungssysteme, sind solche Überraschungen vertretbar. Wenn KI aber eingesetzt werden soll, um kritische Entscheidungen beim autonomen Fahren zu treffen, will man verstehen, wie sie zu diesen Entscheidungen kommt, um mögliche Fehlfunktionen beheben zu können.

Zu diesem Zweck wurden Erklärbarkeits- und Interpretierbarkeitswerkzeuge entwickelt, um insbesondere Wissenschaftlerinnen und Wissenschaftlern sowie Anwendenden auf dem Gebiet des maschinellen Lernens zu einem besseren Verständnis der Funktionsweise von neuronalen Netzen zu verhelfen. Forschende entwickeln Ansätze, die darauf abzielen, die Entscheidungsprozesse von Künstlicher Intelligenz besser zu erklären (Guidotti et al., 2018; Adadi & Berrada, 2018; Arrieta et al., 2020; Samek et al., 2019; Carvalho et al., 2019), indem verschiedene Methoden wie Feature-Wichtigkeitsbewertungen, kontrafaktische Erklärungen oder einflussreiche Trainingsdaten verwendet werden. Allerdings ist bisher unklar, zu welchem Grad diese Methoden in der Praxis angewandt werden. Mehrere interpretierbare Modelle wurden, teils mit großer Resonanz, in der Literatur vorgeschlagen, aber nur wenige Studien untersuchen, ob diese Modelle ihre beabsichtigten Ziele erreichen oder nicht (Wirth et al., 2022), zum Beispiel Menschen dazu zu bringen, den Vorhersagen eines Modells mehr zu folgen, wenn es für sie nützlich ist, dies zu tun, oder ihnen zu ermöglichen, zu erkennen, wenn ein Modell einen Fehler gemacht hat.

Wesentlich für die Wirkung der Erklärung ist dabei eine zielgruppenorientierte, angemessene Komplexität, welche sich aus dem Handlungskontext ergibt. Die in der Einleitung vorgestellten drei Zielgruppen sind auch für die Diskussion um XAI zielführend. Die Literatur legt nahe, dass detaillierte Erklärbarkeitstechniken hauptsächlich von Ingenieurinnen und Ingenieuren sowie Datenwissenschaftlerinnen und Datenwissenschaftlern verwendet werden, um Modelle vor der Bereitstellung zu prüfen oder nach Unfällen den Hergang rekonstruieren zu können. Während ML-Entwicklerinnen und -Entwickler zunehmend solche Erklärbarkeitstechniken als Plausibilitätsprüfungen während des Entwicklungsprozesses verwenden, gibt es immer noch erheblichen Entwicklungsaufwand für Lösungen, um Endbenutzende direkt zu informieren und beispielsweise das Verhalten von Assistenzsystemen nachvollziehbar zu gestalten.

Erklärbarkeit – Zielgerichtete Erklärungen

- Erklärbarkeit unterstützt Vertrauen in ein Produkt, wenn die Erklärungen auf die Ziele abgestimmt sind. Die Ziele können dabei je nach Phase des Produktzyklus unterschiedlich sein: Während der Entwicklung und Zulassung, um beispielsweise Fehler im System leichter entdecken zu können, oder auch um einen beispielsweise im Straßenverkehr geforderten Sicherheitsnachweis angemessen zu untermauern. In dieser Phase ist XAI insbesondere für Unternehmen und deren Entwicklerinnen und Entwickler sowie Institutionen und Behörden für die Prüfung des Verhaltens relevant.

- Während der Anwendung, um sinnvolles Feedback zu geben und das momentane Verhalten interaktiv und nachvollziehbar zu gestalten. Hier sollten die Erklärungen hauptsächlich der Nachvollziehbarkeit für Nutzerinnen und Nutzer dienen, beispielsweise im Falle eines Eingreifens von Assistenzsystemen.

- Nach der Anwendung, um eine Fehlfunktion analysieren zu können, beispielsweise zum Zweck einer iterativen Verbesserungsschleife oder zur Klärung einer Haftungsfrage im Falle eines Verkehrsunfalls. Hier ist XAI primär für Institutionen und Behörden wichtig, um Risikosituationen und Unfälle nachvollziehen zu können.

Diskussion „Wie viel Erklärung ist nötig?"

Ein möglicher Ansatz betrachtet KI als ein nicht durchgängig erklärbares System, das durch rigorose Praktiken in Software- und Sicherheits-Engineering entwickelt und – ähnlich dem Menschen in einer Führerscheinprüfung – in umfassenden und ausgewählten Szenarien analysiert und getestet und bei Erfolg als für diese Szenarien geeignet eingestuft wird. Vergleichbar mit einem menschlichen Nutzenden wird einer erfolgreich getesteten KI vertraut, dass sie den zu erwartenden Situationen gewachsen ist. Dabei wird davon ausgegangen, dass eine KI genauso wie andere Software zu behandeln ist. So ist es für die Entwicklung klassischer Softwaresysteme üblich, Prozesse zu definieren, welche spezifische Sicherheitsbedenken durch konkrete Maßnahmen in der Softwareentwicklung adressieren. Entsprechend dem praktizierten Vorgehen widmet sich eine sichere ML-Softwareentwicklung dem Ziel, konkrete Maßnahmen zu definieren (zum Beispiel Robustheit, Datenqualität, Unsicherheitsberechnung, Runtime-Monitoring, Out-of-Distribution), um auch für ML-Software ein Vertrauensmaß zu gewinnen, welches den Anforderungen entspricht.

Bei dieser Betrachtungsweise werden die bisher klassischerweise benutzten Möglichkeiten zur Vertrauensbildung wie das Testen wieder in den Vordergrund gerückt. Lediglich in dem Fall, dass ein lernendes System eine Fehlentscheidung getroffen hat, sind Erklärbarkeit und Nachvollziehbarkeit hilfreich, um so das betreffende System zu optimieren, diese Situation zu lernen und zukünftig ein angemessenes Verhalten zu zeigen. Solche unbekannten Situationen oder Fehlentscheidungen sollten nicht erst bei Unfällen herangezogen, sondern frühzeitig durch einen intelligenten Systemmonitor als sogenannte „Corner Cases" detektiert und analysiert werden (Hesse, Peylo et al., 2021).

In diesem Kontext kann argumentiert werden, dass Erklärbarkeit in Form von XAI zwar hilfreich, aber nicht durchgängig notwendig ist, um vertrauenswürdige Systeme mit technisch nachvollziehbarem Verhalten zu realisieren. Auch anderen technischen Systemen wird von den Anwendenden Vertrauen entgegengebracht, ohne dass diese alle Funktionen im Detail nachvollziehen können. Wichtig ist lediglich, dass sich die Systeme

den Erwartungen entsprechend verhalten – die technische Funktionsweise der Systeme ist dagegen häufig sekundär. Die Forderung, für KI andere Regeln als für gewöhnliche Software anzuwenden, sollte mit dem damit verbundenen Aufwand in Relation gesetzt werden. Obgleich es sich bei KI um eine Technologie handelt, die anders funktioniert und konstruiert ist als herkömmliche, bereits etablierte Technologien, bleibt jene im Kern doch ein technisches System, das entsprechend etablierten Standards geprüft und getestet werden kann. Allerdings ist es notwendig, die klassischen Standards und Methoden so zu interpretieren und anzupassen, dass sie auf KI anwendbar sind, wie beispielsweise in der ISO/PAS 8800 im sicherheitskritischen Umfeld vorgegangen wird (International Organization for Standardization, ISO/PAS 8800).

5. Gestaltungsoptionen

Vertrauen ist ein zentrales Element für die Akzeptanz neuer, KI-basierter Technologien. Es wird vor allem aufgebaut durch nützliche Anwendungen und Partizipation. Um Partizipation und einen aufgeklärten, gesamtgesellschaftlichen Diskurs zu ermöglichen, ist Verständnis und Kommunikation der eingesetzten Verfahren, Architekturen, Datenakquisition und beteiligten Akteure des KI-Systems zwar wichtig für die Zulassung, aber nicht immer für die Anwendenden relevant. Stattdessen sollte für unterschiedliche Zielgruppen eine jeweils angepasste Informationstiefe angestrebt werden. Im Folgenden werden für die drei in Abbildung 1 beschriebenen Zielgruppen mögliche Gestaltungsoptionen beschrieben.

Anwendende und Gesellschaft

Damit das Potenzial KI-basierter Mobilitätsanwendungen ausgeschöpft werden kann, bedarf es des Vertrauens in die Anwendungen und in deren Vorteile, insbesondere seitens der Nutzerinnen und Nutzer.

- Das Potenzial neuer Technologien sollte fair beurteilt werden. Oftmals neigen Menschen dazu, Vertrauen in Bekanntes zu setzen, selbst wenn objektiv betrachtet Neues die bessere Wahl wäre. Dieses Phänomen lässt sich beispielsweise gut anhand eines Vergleichs des subjektiven Sicherheitsgefühls bei einem Linienflug und einer Fahrradfahrt illustrieren. Während die vertraute Art der Fortbewegung mit dem Fahrrad zumeist als sicherer wahrgenommen wird, ist bei statistischer Betrachtung (Statista, 2023; Statista, 2024) das Flugzeug mit großem Abstand das sicherere Verkehrsmittel. Es gilt daher, ein ausgeglichenes Verständnis für die Risiken und Chancen von Innovationen anzustreben.

- Nutzende sollten prüfen, ob sich die KI-Systeme auf ihre Präferenzen anpassen lassen und ob erlernte Verhaltensweisen, damit generell die Bedingungen für das Verhalten des Systems für die Nutzenden, nachvollziehbar gestaltet sind.

- Wichtig ist eine aktive Auseinandersetzung mit den neuen Technologien und ein aufgeklärter Umgang mit ihnen. Dabei gilt es, sich bewusst zu sein, dass die Fähigkeiten der Technologie aus Marketinggründen übertrieben dargestellt werden können, während in der Medienbranche oft Mechanismen wirken, die Schlagzeilen gegenüber objektiven Analysen bevorzugen. Ein solcher Umgang erlaubt eine bewusste Entscheidung für den Einsatz und zur Nutzung von KI-Systemen.

- Eine aktive Beteiligung an einem aufgeklärten, gesamtgesellschaftlichen Diskurs zu Chancen und Risiken von KI in der Mobilität sollte gesucht werden. In diesem Diskurs darf die technische Perspektive nicht außer Acht gelassen werden, um die Ergebnisse des Diskurses auch technisch umsetzen zu können.

Unternehmen und Anbieter

Unternehmen und Anbieter sollten technische Systeme in ihrer Gänze transparent und nachvollziehbar gestalten und die Nutzerinnen und Nutzer beispielsweise über Umfang und Verwendungszweck gesammelter Daten informieren. Wie in der KI-Strategie der Bundesregierung (Bundesministerium für Wirtschaft und

Energie, 2018) formuliert, sollte dabei stets der Nutzen für die Anwendenden im Fokus stehen. Hier ist auf eine angemessene Komplexität des Informationsgehalts zu achten – eine Überforderung der Nutzenden mit allen verfügbaren Informationen ist somit nicht zielführend.

- Ein guter Ausgangspunkt ist es, Anwendungsfälle zu identifizieren und umzusetzen, die zu Beginn bereits einen sicheren Nutzen erlauben, auch wenn dieser Nutzen nur einen Teilbereich eines größeren Zielnutzens erfüllt. Dies ermöglicht von Anfang an ein positives und vertrauensbildendes Nutzererlebnis.

- Anbieter sollten ihren Nutzenden verwendete Daten sowie die Priorisierung von Entscheidungen transparent machen und somit auch ein Verständnis der eingesetzten Verfahren ermöglichen.

- Systeme sollten individuell einstellbar sein, um verschiedene Präferenzen der Interaktion mit technischen Systemen zu ermöglichen, gleichzeitig aber den Raum des regelkonformen Verhaltens niemals verlassen können. Hinzu kommt, dass Korrekturen vom System wahrgenommen und als Präferenz gelernt werden und dass im Weiteren diese gelernten Verhaltensweisen, damit generell die Bedingungen für das Verhalten des Systems für die Nutzenden, transparent bleiben.

- Die Fähigkeiten der Systeme sollten realistisch dargestellt werden, anstatt durch Marketingversprechen die Kundinnen und Kunden zu überhöhten Ansprüchen an die Technologie zu bringen. Dies ist wichtig, damit sich die Systeme den Erwartungen entsprechend beweisen können, statt die Nutzenden zu enttäuschen und so das Vertrauen zu mindern.

- Bei der Entwicklung KI-basierter Systeme müssen kulturelle und ethische Unterschiede in verschiedenen Regionen der Welt berücksichtigt werden. Dies kann gelingen, wenn nicht allein die Technik im Mittelpunkt steht, sondern deren Integration und Anwendung im Mobilitätssystem.

Institutionen und Politik

Institutionen und Politik sollten sicherstellen, dass Regeln, aber auch Versprechen eingehalten werden. Sie sollten auch die in der Mobilität unvermeidbaren Zielkonflikte möglichst optimal ausbalancieren.

- Institutionen und Politik sollten einen offenen und aufgeklärten gesellschaftlichen Diskurs über KI in der Mobilität initiieren und fördern.

- Die schrittweise Integration von Reallaboren in den öffentlichen Raum ermöglicht eine praxisnahe Erforschung neuer Technologien. Ein zentraler Aspekt dabei ist die Gewährleistung der Sicherheit.

- Für Unternehmen ist es wichtig, klare Regeln und rechtliche Rahmenbedingungen zu definieren. Gleichzeitig muss sichergestellt werden, dass Regeln, aber auch Versprechen bei Zulassung und Inverkehrbringung eingehalten werden.

- Die in der Mobilität unvermeidbaren Zielkonflikte sollten möglichst transparent gemacht und möglichst optimal ausbalanciert werden.

Literatur

Adadi, A. & Berrada, M. (2018): Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Arrieta, B. et al. (2020): Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bundesministerium der Justiz (n.d.): Straßenverkehrs-Ordnung (StVO) § 1 Grundregeln.

Bundesministerium für Wirtschaft und Energie (2018): KI-Strategie Deutschland. Online unter: https://www.ki-strategie-deutschland.de/home.html

Carvalho, V., Pereira, E. M. & Cardoso, J. S. (2019): Machine learning interpretability: A survey on methods and metrics. Electronics, 8(8), Article 832. https://doi.org/10.3390/electronics8080832

Choi, J. K. & Ji, Y. G. (2015): Investigating the importance of trust on adopting an autonomous vehicle. International Journal of Human–Computer Interaction, 31(10), 692–702. https://doi.org/10.1080/10447318.2015.1070549

Clarke, Y. D. (2019): H.R.2231 – Algorithmic Accountability Act of 2019. Retrieved from https://www.congress.gov/bill/116th-congress/house-bill/2231

Dahlbäck, N., Jönsson, A. & Ahrenberg, L. (1993): Wizard of Oz studies – why and how. Knowledge-Based Systems, 6(4), 258–266. https://doi.org/10.1016/0950-7051(93)90017-N

Davis, F. (1985): A technology acceptance model for empirically testing new end-user information systems – theory and results. PhD thesis, Massachusetts Inst. of Technology.

DEKRA (2023): Verkehrssicherheitsreport 2023 „Technik und Mensch".

DKE und DIN (2022): Ethik und Künstliche Intelligenz: Was können technische Normen und Standards leisten? Whitepaper. Retrieved from https://www.din.de/resource/blob/754724/00dcbccc21399e13872b2b6120369e74/whitepaper-ki-ethikaspekte-data.pdf

Duden: Vertrauen, das: Online unter: https://www.duden.de/rechtschreibung/Vertrauen

European Commission (2019): Ethics guidelines for trustworthy AI. Online unter: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Parliament (2024): Artificial Intelligence Act. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=EP%3AP9_TA%282024%290138

European Parliament and Council of the European Union (2016): Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Online unter: https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=celex%3A32016R0679

Farke, F. M. et al. (2021): Are privacy dashboards good for end users? Evaluating user perceptions and reactions to Google's My Activity. 30th USENIX Security Symposium (USENIX Security 21).

Foot, P. (1978): The problem of abortion and the doctrine of the double effect. In Virtues and vices (pp. 19–32). Oxford: Basil Blackwell. (Original work published 1967 in the Oxford Review, Number 5)

Guidotti, R. et al. (2018): A survey of methods for explaining black box models. ACM Computing Surveys (CSUR), 51(5), Article 93. https://doi.org/10.1145/3236009

Hesse, T., Peylo, C. et al. (2021): Potenziale für industrieübergreifendes Flottenlernen – KI-Mobilitätsdatenplattform zur Risikominimierung des automatisierten Fahrens. Whitepaper aus der Plattform Lernende Systeme. München.

Hochrangige Expertengruppe für künstliche Intelligenz (2019): Ethics guidelines for trustworthy AI. Online unter: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Hungund, A., Pai, G. & Pradhan, A. K. (2021): Systematic review of research on driver distraction in the context of advanced driver assistance systems. Transportation Research Record: Journal of the Transportation Research Board, 2675(9), 036119812110041. https://doi.org/10.1177/03611981211004129

International Organization for Standardization (in development): ISO PAS 8800: Road Vehicles – Safety and artificial intelligence.

Jelinski, L., Etzrodt, K. & Engesser, S. (2021): Undifferentiated optimism and scandalized accidents: The media coverage of autonomous driving in Germany. JCOM, 20(04), A02. https://doi.org/10.22323/2.20040202

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012): ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems 25 (pp. 1097–1105). Retrieved from https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

Lyu, N., Duan, Z., Ma, C. & Wu, C. (2020): Safety margins – A novel approach from risk homeostasis theory for evaluating the impact of advanced driver assistance systems on driving behavior in near-crash events. Journal of Safety Research, 54, 179–5846. https://doi.org/10.1080/15472450.2020.1795846

Molnar, C. (2020): Interpretable machine learning: A guide for making Black Box models explainable. Eigenpublikation.

OECD/ITF (2015): Automated and Autonomous Driving: Regulation under Uncertainty. Abgerufen: 22.11.23 von https://www.oecd-ilibrary.org/docserver/5jlwvzdfk640-en.pdf?expires=1721820225&id=id&accname=guest&checksum=232FA45DB96A58C129FD9E58F792C46E

Plattform Lernende Systeme (2022): Fiktives Gerichtsverfahren: Wer haftet für Schäden durch autonome Fahrzeuge? https://www.plattform-lernende-systeme.de/aktuelles-newsreader/fiktives-gerichtsverfahren-wer-haftet-fuer-schaeden-durch-autonome-fahrzeuge.html

Plattform Lernende Systeme (2024): KI Kompakt: AI Act der Europäischen Union: Regeln für vertrauenswürdige KI (Publikationsreihe). Online unter: https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/KI_Kompakt/KI_Kompakt_AI_Act_Plattform_Lernende_Systeme_2024.pdf

Salem, M. et al. (2015): Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI '15) (pp. 141–148). Association for Computing Machinery. https://doi.org/10.1145/2696454.2696497

Samek, W. et al. (2019): Explainable AI: Interpreting, explaining, and visualizing deep learning (Vol. 11700). Springer Nature.

Schlag, B. (2016): Automatisiertes Fahren im Straßenverkehr – Offene Fragen aus Sicht der Psychologie. Zeitschrift für Verkehrssicherheit, 62.

Statista (2023): Verkehrsunfälle: Statista Dossier. Statista. https://de.statista.com/statistik/studie/id/6890/dokument/verkehrs-unfaelle-statista-dossier/

Statista (2024): Sicherheit in der Luftfahrt: Statista Dossier. Statista. Online unter: https://de.statista.com/statistik/studie/id/30084/dokument/sicherheit-in-der-luftfahrt-statista-dossier/

U.S. Department of Transportation, Office of the Secretary of Transportation (2016): Enhanced FAA oversight could reduce hazards associated with increased use of flight deck automation. Report Number: AV-2016-013. Abgerufen: 23.12.23 von https://www.oig.dot.gov/sites/default/files/FAA%20Flight%20Decek%20Automation_Final%20Report%5E1-7-16.pdf

Wang, N., Pynadath, D. V., & Hill, S. G. (2016a): Trust calibration within a human-robot team: Comparing automatically generated explanations. In The Eleventh ACM/IEEE International Conference on Human Robot Interaction (HRI '16) (pp. 109–116). IEEE Press.

Wang, N., Pynadath, D. V. & Hill, S. G. (2016b): The impact of POMDP-generated explanations on trust and performance in human-robot teams. In Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems (AAMAS '16) (pp. 997–1005). International Foundation for Autonomous Agents and Multiagent Systems.

Weiss, A. et al. (2009): User experience evaluation with a Wizard of Oz approach: Technical and methodological considerations. In 9th IEEE-RAS International Conference on Humanoid Robots (pp. 303–308). https://doi.org/10.1109/ICHR.2009.5379559

Wirth, C., Schmid, U. & Voget, S. (2022): Humanzentrierte künstliche Intelligenz: erklärendes interaktives maschinelles Lernen für Effizienzsteigerung von Parmetierungsaufgaben. In E. A. Hartmann (Ed.), Digitalisierung souverän gestalten II (pp. 80–92). Springer Vieweg. ISBN 978-3-662-64407-2.

Über dieses Whitepaper

Die Autoren des Whitepapers sind Mitglieder der Arbeitsgruppen *Mobilität und intelligente Verkehrssysteme* sowie *IT-Sicherheit, Privacy, Recht und Ethik* der Plattform Lernende Systeme.

Als eine von insgesamt sieben Arbeitsgruppen untersucht die Arbeitsgruppe *Mobilität und intelligente Verkehrssysteme*, wie Lernende Systeme unsere Mobilitätsstrukturen verändern und welche Eigenschaften sie haben müssen, um den größten Nutzen für das Individuum und die Gesellschaft zu erzielen. Die Arbeitsgruppe hinterfragt, wie Infrastrukturen und Systemarchitekturen im Mobilitätssektor weiterentwickelt werden müssen, um Lernende Systeme darin sinnvoll zu integrieren. Die Arbeitsgruppe *IT-Sicherheit, Privacy, Recht und Ethik* thematisiert Fragen zur Sicherheit (Security), Zuverlässigkeit (Safety) und zum Umgang mit Privatheit (Privacy) bei der Entwicklung und Anwendung von Lernenden Systemen. Sie analysiert zudem damit verbundene rechtliche sowie ethische Anforderungen und steht in engem Austausch mit allen weiteren Arbeitsgruppen der Plattform Lernende Systeme.

Mitglieder der Arbeitsgruppe Mobilität und intelligente Verkehrssysteme
Dr. Claus Bahlmann, Siemens Mobility, AG Leitung
Dr.-Ing. Tobias Hesse, Deutsches Zentrum für Luft- und Raumfahrt e. V. (DLR), Institut für Verkehrssystemtechnik
Prof. Dr.-Ing. Fabian Behrendt, Fraunhofer-Institut für Fabrikbetrieb und -automatisierung IFF
Dr. Rudolf Felix, PSI FLS Fuzzy Logik & Neuro Systeme GmbH
Prof. Dr.-Ing. Axel Hahn, Deutsches Zentrum für Luft- und Raumfahrt e. V. (DLR), Institut Systems Engineering für zukünftige Mobilität
Dr.-Ing. Sören Kerner, Fraunhofer-Institut für Materialfluss und Logistik IML
Sascha Ott, Karlsruher Institut für Technologie (KIT)/Institut für Produktentwicklung (IPEK)
Dr.-Ing. Ilja Radusch, Fraunhofer FOKUS/Daimler Center for Automotive IT Innovations
Dr. Peter Schlicht, CARIAD SE
Prof. Dr.-Ing. Philipp Slusallek, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)
Dr. Stefan Voget, Continental AG
Prof. Dr.-Ing. J. Marius Zöllner, Karlsruher Institut für Technologie (KIT)

Mitglieder der Unterarbeitsgruppe Recht und Ethik
Prof. Dr. Michael Decker, Karlsruher Institut für Technologie (KIT)/Bereichsleitung „Informatik, Wirtschaft und Gesellschaft"
Prof. Dr. Armin Grunwald, Karlsruher Institut für Technologie (KIT)/Institut für Technikfolgenabschätzung und Systemanalyse (ITAS)

Ehemalige Mitglieder der Arbeitsgruppe Mobilität und intelligente Verkehrssysteme
Dr. Christoph Peylo, Bosch Center for Artificial Intelligence (BCAI)

Redaktion
Patrick Bollgrün, Geschäftsstelle der Plattform Lernende Systeme
Christine Wirth, Geschäftsstelle der Plattform Lernende Systeme

Impressum

Herausgeber
Lernende Systeme –
Die Plattform für Künstliche Intelligenz
Geschäftsstelle | c/o acatech
Karolinenplatz 4 | 80333 München
www.plattform-lernende-systeme.de

Gestaltung und Produktion
PRpetuum GmbH, München

Stand
September 2024

Bildnachweis
AdobeStock/Techtility Design

Empfohlene Zitierweise
Bahlmann, C., Felix, R., Hahn, A. et al. (2024): Vertrauen in KI-basierte Mobilität. Technologische und ethische Aspekte. Whitepaper aus der Plattform Lernende Systeme, München.
DOI: https://doi.org/10.48669/pls_2024-7

Bei Fragen oder Anmerkungen zu dieser Publikation kontaktieren Sie bitte Dr. Thomas Schmidt (Leiter der Geschäftsstelle):
kontakt@plattform-lernende-systeme.de

Über die Plattform Lernende Systeme

Die Plattform Lernende Systeme ist ein Netzwerk von Expertinnen und Experten zum Thema Künstliche Intelligenz (KI). Sie bündelt vorhandenes Fachwissen und fördert als unabhängiger Makler den interdisziplinären Austausch und gesellschaftlichen Dialog. Die knapp 200 Mitglieder aus Wissenschaft, Wirtschaft und Gesellschaft entwickeln in Arbeitsgruppen Positionen zu Chancen und Herausforderungen von KI und benennen Handlungsoptionen für ihre verantwortliche Gestaltung. Damit unterstützen sie den Weg Deutschlands zu einem führenden Anbieter von vertrauenswürdiger KI sowie den Einsatz der Schlüsseltechnologie in Wirtschaft und Gesellschaft. Die Plattform Lernende Systeme wurde 2017 vom Bundesministerium für Bildung und Forschung (BMBF) auf Anregung des Hightech-Forums und acatech – Deutsche Akademie der Technikwissenschaften gegründet und wird von einem Lenkungskreis gesteuert.