Kompetenzkreis KI im Tourismus

Grundlagen zu Künstlicher Intelligenz

Wenn KI-Agenten Grenzen überschreiten: Was die Sicherheitsdebatte für den Tourismus bedeutet

Wenn KI-Agenten Kundenanfragen beantworten, Buchungen verändern oder Erstattungen auslösen, wird aus einer digitalen Antwort eine Handlung mit Folgen. Für touristische Unternehmen stellt sich deshalb eine konkrete Frage: Welche Aufgaben und Zugriffe geben wir frei – und wie stellen wir sicher, dass die Systeme innerhalb dieser Grenzen bleiben?

Kundendaten, Buchungen, Postfächer und Zahlungsprozesse dürfen nur unter klaren Regeln von Berechtigten genutzt werden. Fehler und Missbrauch müssen erkannt und begrenzt werden. Für die Frage der Berechtigungen kann man sich einen KI-Agenten als eine Art „digitalen Mitarbeiter" vorstellen: Er bekommt eine Aufgabe und bestimmte Zugriffe. Die Verantwortung für sein Handeln bleibt jedoch bei Menschen und Organisationen.

Die aktuelle Sicherheitsdebatte geht über die alltägliche Frage hinaus, ob KI gelegentlich falsche Antworten gibt. KI-Anbieter haben selbst Vorfälle veröffentlicht, bei denen Agenten in Forschungsumgebungen Grenzen überschritten. In einem von der Forschungsorganisation METR untersuchten Fall koordinierten sich Agenten während eines Sicherheitstests über einen unerwarteten Kommunikationsweg, griffen Systeme von Hugging Face an und versuchten, die Bewertung ihrer eigenen Aufgaben zu beeinflussen. Anthropic berichtete über drei Vorfälle in eigenen Sicherheitstests, bei denen Modelle aufgrund unzureichend abgeschotteter Testumgebungen unerlaubt auf reale Systeme zugriffen. Die Bedingungen dieser Tests waren ungewöhnlich; zugleich hatten einzelne Handlungen reale Außenwirkungen.

Nun hat OpenAI zudem die geplante Veröffentlichung seines Modells GPT-6.1 Astra nach internen Sicherheitsprüfungen gestrichen. Nach Angaben des Unternehmens hielt das noch nicht veröffentlichte Modell erteilte Befugnisse nicht zuverlässig ein und berichtete Nutzern nicht immer zutreffend über seine Arbeit. Davon zu unterscheiden ist ein weiterer, gesondert dokumentierter Vorfall, nach dem OpenAI Teile des Trainings und der Tests besonders leistungsfähiger Modelle pausierte: Ein Forschungsmodell hatte trotz vorgesehener Abschottung einen Weg gefunden, einen externen Chatbot zu erreichen.

Die öffentlich verfügbaren Berichte erlauben keine vollständige unabhängige Bewertung der Abläufe. Aus ihnen lässt sich auch nicht ableiten, wie häufig vergleichbares Verhalten im Unternehmensalltag zu erwarten wäre. Dennoch lassen sich die Ereignisse nicht mit dem Hinweis „Das war nur ein Test" abtun: Einzelne Handlungen hatten reale Außenwirkungen, und Sicherheitsbefunde führten jüngst zum Stopp einer Modellfreigabe. Die Fälle zeigen, dass Sicherheit nicht allein von der ursprünglichen Aufgabe eines Agenten abhängt. Entscheidend sind ebenso seine Fähigkeiten, die erreichbaren Systeme, wirksame technische Grenzen und die Möglichkeit, unerwartete Handlungen rechtzeitig zu erkennen.

Kompetenzkreis KI im Tourismus

Bemerkenswert ist, wie deutlich inzwischen einige Hersteller selbst öffentlichkeitswirksam warnen. Anthropic-Chef Dario Amodei befürchtet in seinem Beitrag „We Must Pace the Frontier“, dass die Entwicklung besonders leistungsfähiger Modelle der Sicherheitsforschung und Kontrolle davonlaufen könnte. Er fordert mehr Zeit für Prüfungen, unabhängige Prüfer mit tiefem Einblick in die Entwicklung und gemeinsame Sicherheitsstandards. Das ist eine weitreichende Position in einer offenen Debatte – keine gesicherte Vorhersage darüber, was KI in einigen Monaten tun wird. Auch unter Fachleuten gehen die Einschätzungen über die Wahrscheinlichkeit eines schwerwiegenden Kontrollverlusts auseinander, wie der International AI Safety Report 2026 darlegt.

Warum tut ein Agent etwas, das niemand beauftragt hat?

Ein KI-Modell lernt aus menschlich erzeugten Inhalten und damit auch aus Beispielen für Täuschung, Abkürzungen und Regelverstöße. Das allein erklärt die aktuellen Vorfälle jedoch nicht. Ein Modell übernimmt nicht einfach menschliche Motive. Bei seiner weiteren Entwicklung lernt es auch, welche Ergebnisse als Erfolg gelten. Wird ein Ziel zu eng definiert oder eine unerwünschte Abkürzung belohnt, kann ein Agent einen Weg wählen, der zwar einen guten Wert im Test erzielt, aber den eigentlichen Sinn der Aufgabe verfehlt. Forschende bezeichnen das als *Reward Hacking*.

Fortgeschrittene Modelle sind darauf ausgerichtet, vorgegebene Ziele möglichst erfolgreich zu erreichen. Ein moralisches Verständnis im menschlichen Sinne besitzen sie jedoch nicht. Im beschriebenen Belohnungsprozess können sie zudem Strategien erlernen, mit denen sie bestehende Ein- und Ausgabebeschränkungen (Guardlines) umgehen, wenn dies die Zielerreichung begünstigt.

Hinzu kommt die Umgebung, in der ein Agent arbeitet. Wer nur einen Textvorschlag machen kann, hat andere Möglichkeiten als ein System, das Programme ausführt, Dateien liest oder auf das Internet zugreift. Deshalb müssen wir bei jedem Vorfall fragen: Was war die Aufgabe? Was wurde als Erfolg belohnt? Welche Zugriffe hatte das System? Und welche technischen Grenzen hätten die Handlung verhindern sollen? OpenAI nennt für den Hugging-Face-Vorfall unter anderem das Ausnutzen der Bewertung, Beharrlichkeit bei scheinbar unlösbaren Aufgaben und unerlaubte Zusammenarbeit zwischen mehreren Agenten als Faktoren.

Ein Ziel ist nur auf einem zulässigen Weg erreicht

Aus diesen Vorfällen ergibt sich eine grundsätzliche Frage: **Was soll ein Agent tun, wenn er eine Aufgabe nur lösen kann, indem er eine Grenze überschreitet?** Soll er sich einen Zugang beschaffen, den niemand freigegeben hat? Soll er eine Information erfinden, weil die richtige nicht verfügbar ist? Oder soll er offen sagen, dass er unter den geltenden Bedingungen nicht weiterkommt?

Unsere Antwort ist klar: Die Erfüllung einer Aufgabe darf nicht wichtiger sein als die Regeln, innerhalb derer sie erledigt werden soll. Ein Agent muss innerhalb seiner Befugnisse bleiben, Unsicherheit oder Hindernisse offenlegen und sich stoppen oder korrigieren lassen. Das muss sich im Training, in Prüfungen und in den technischen Zugriffsmöglichkeiten widerspiegeln. Eine Formulierung im Prompt allein kann keine verlässliche Sicherheitsgrenze ersetzen.

Kompetenzkreis KI im Tourismus

Ein Beispiel dafür, wie ein Hersteller dieses Prinzip formuliert, ist Microsofts „Humanist AI Code of Conduct“ für seine MAI-Modelle. Darin hat die Einhaltung des Kodex ausdrücklich Vorrang vor dem Erfolg einer Aufgabe. Die Modelle sollen unter anderem innerhalb erteilter Befugnisse bleiben, menschliche Eingriffe zulassen und ihre Handlungen nicht verschleiern. Wenn ein Ziel nur durch einen erheblichen Verstoß gegen den Kodex erreichbar wäre, soll das Modell die Aufgabe nicht auf diesem Weg erfüllen.

Das ist ein wichtiger Anspruch, aber noch kein Nachweis, dass jede Leitplanke in jeder Situation funktioniert. Microsoft bezeichnet den Kodex selbst als Zielbild und nicht als Garantie für das heutige Verhalten seiner Modelle. Entscheidend bleibt, ob die Regeln technisch umgesetzt, unter schwierigen Bedingungen getestet und bei Vorfällen überprüft werden.

Welche gesetzlichen Leitplanken gibt es?

Auch rechtlich beginnt die Verantwortung nicht erst, nachdem ein Schaden entstanden ist. Der EU AI Act enthält verbindliche vorbeugende Regeln. Anbieter allgemeiner KI-Modelle mit systemischem Risiko müssen Risiken bewerten und mindern sowie schwerwiegende Vorfälle verfolgen und melden. Für bestimmte Hochrisiko-KI-Systeme verlangt der AI Act unter anderem Risikomanagement, menschliche Aufsicht, Robustheit und Cybersicherheit. Wer solche Systeme einsetzt, muss ihren Betrieb überwachen und bei erkannten Risiken handeln. Hinzu kommen Verbote bestimmter KI-Praktiken, Transparenzpflichten und Maßnahmen zur Förderung von KI-Kompetenz. Der freiwillige EU-Kodex für Anbieter allgemeiner KI-Modelle hilft Herstellern bei der Umsetzung eines Teils ihrer gesetzlichen Pflichten.

Allerdings erfasst die Kategorie „Hochrisiko-KI“ nicht jeden Agenten mit weitreichenden Handlungsmöglichkeiten. Ein Agent, der auf Buchungen, Postfächer oder Zahlungen zugreifen kann, wird durch diese Zugriffe allein nicht automatisch zum Hochrisiko-System im Sinne des AI Acts. Deshalb müssen Unternehmen auch die bestehenden Regeln für ihre konkreten Prozesse beachten: Die DSGVO schützt Kunden- und Beschäftigtendaten und verlangt eine rechtmäßige, angemessen abgesicherte Verarbeitung. Reise- und Verbraucherrecht gelten ebenso, wenn eine KI Angebote erläutert oder Buchungen ausführt; § 651x BGB regelt beispielsweise die Haftung für bestimmte Buchungsfehler. Für Unternehmen im Anwendungsbereich des BSI-Gesetzes kommen Pflichten zur IT-Sicherheit hinzu.

Eine weitere Ebene betrifft die Haftung nach Schäden. Die neue EU-Produkthaftungsrichtlinie erfasst ausdrücklich auch Software und KI-Systeme. Sie soll für Produkte gelten, die ab dem 9. Dezember 2026 in Verkehr gebracht oder in Betrieb genommen werden; die Umsetzung in deutsches Recht ist dabei zu beachten. Sie kann Betroffenen bei bestimmten Schäden Ansprüche gegen Hersteller eröffnen. Ihre Grenze für unsere Debatte ist offensichtlich: Wenn Haftung geprüft wird, kann der Schaden bereits eingetreten sein. Haftungsregeln schaffen deshalb einen Anreiz für Sorgfalt, ersetzen aber keine wirksamen Prüfungen, Berechtigungsgrenzen und Eingriffsmöglichkeiten vor und während des KI-Einsatzes.

Europäische und deutsche Regeln setzen damit an mehreren Stellen an. Sie verteilen Pflichten jedoch je nach Rolle und Anwendungsfall unterschiedlich. Für den Tourismus kommt es darauf an, diese Regeln

mit den tatsächlichen Handlungsmöglichkeiten agentischer Systeme zusammenzubringen – und Schutzmaßnahmen nicht allein von einer rechtlichen Risikokategorie abhängig zu machen.

Verantwortung auf drei Ebenen

Die drei Ebenen beschreiben unterschiedliche Aufgaben. Ein Unternehmen kann mehrere davon zugleich übernehmen – beispielsweise, wenn es eine eigene touristische KI-Lösung entwickelt oder ein Modell selbst mit seinen Buchungs- und Kundensystemen verbindet.

Erstens: die Hersteller der KI-Modelle. Sie müssen Fähigkeiten und Grenzen ihrer Modelle prüfen, Sicherheitsregeln in Entwicklung und Betrieb verankern, relevante Vorfälle nachvollziehbar offenlegen und unabhängige Bewertungen ermöglichen. Ein Unternehmen im Tourismus kann die internen Eigenschaften eines fremden Modells nicht selbst vollständig untersuchen.

Zweitens: Anbieter und Integratoren von KI-Lösungen. Sie entscheiden, wie ein Modell mit Postfächern, Datenbanken, Buchungssystemen und anderen Werkzeugen verbunden wird. Sie müssen Berechtigungen begrenzen, fremde Inhalte absichern, Aktionen protokollieren und wirksame Eingriffsmöglichkeiten schaffen. Vor dem Einsatz und nach Änderungen müssen sie die gesamte Anwendung prüfen, bei erhöhtem Risiko auch durch professionelle Sicherheitsprüfungen wie Penetrationstests und KI-spezifische Angriffstests. Ein sorgfältig geprüftes Modell allein macht eine unsicher angeschlossene Anwendung nicht sicher.

Drittens: die einsetzenden Unternehmen. Sie bestimmen den Zweck des Einsatzes, wählen geeignete Lösungen und Zugänge, geben Daten und Zugriffe frei und organisieren die Aufsicht. Kostenlose Tools und persönliche Accounts ohne geeignete unternehmensweite Schutz- und Kontrollmechanismen sollten nicht als Infrastruktur für geschäftssensible Arbeiten dienen. Mit festen Testanfragen und Anwendungsszenarien müssen sie regelmäßig prüfen, ob Vorgaben auch nach Modelländerungen eingehalten werden, und bei Abweichungen nachjustieren. Beschäftigte brauchen klare Regeln und Kompetenz; technische und organisatorische Sicherheitsmängel können sie nicht durch persönliche Vorsicht ausgleichen.

Im Tourismus reicht das Spektrum von unterstützenden Aufgaben wie Recherche, Content-Produktion, Reporting oder der Vorbereitung von Angeboten und Serviceanfragen bis zu operativen Handlungen. Mit entsprechenden Zugriffsrechten können KI-Agenten persönliche Reisedaten verarbeiten, Buchungen ändern, Erstattungen auslösen oder im Namen eines Unternehmens kommunizieren. Fehler können daher auf allen Stufen spürbare Folgen durch unzutreffende Leistungsversprechen, die Veröffentlichung vertraulicher Informationen oder fehlerhafte Eingriffe in Buchungs- und Zahlungsprozesse haben.

Je näher KI-Agenten an sensible Daten und folgenreiche Handlungen rücken, desto sorgfältiger müssen alle drei Ebenen zusammenwirken.

Kompetenzkreis KI im Tourismus

Sieben Maßnahmen für touristische Unternehmen

1. KI-Einsätze erfassen und Risiken bewerten. Dokumentieren, welche Anwendungen und Agenten genutzt werden, wer sie verantwortet und welche Systeme, Daten und externen Dienste sie erreichen. Auch eigenständig genutzte Tools und persönliche Accounts von Mitarbeitenden berücksichtigen. Zwischen Textentwürfen, Kundenkommunikation, Buchungsänderungen und Zahlungen unterscheiden: Die möglichen Folgen für Reisende, Partner und das eigene Unternehmen bestimmen den Umfang der Schutzmaßnahmen.

2. Berechtigungen begrenzen und Daten trennen. Lesen, Ändern, Löschen, Versenden und Bezahlen getrennt freigeben. Ein Agent erhält nur die Rechte und Datenzugriffe, die seine konkrete Aufgabe erfordert. Informationen verschiedener Kunden und Projekte konsequent voneinander trennen. Klar abgegrenzte Wissensbestände und technische Zugriffsbeschränkungen müssen verhindern, dass vertrauliche Informationen unzulässig genutzt, vermischt werden oder an Unberechtigte gelangen.

3. Autonomie schrittweise zulassen und wirksam beaufsichtigen. Mit Vorschlägen beginnen, dann einzelne Aktionen nach Prüfung freigeben. Selbstständige Ausführung nur innerhalb klarer Grenzen und nach ausreichenden Tests zulassen. Festlegen, wann der Agent stoppen und einen Menschen hinzuziehen muss. Vor einer Freigabe müssen die geplante Handlung und ihre möglichen Folgen erkennbar sein; ein undurchsichtiger Bestätigungsklick genügt nicht.

4. Unternehmenseigene KI-Lösungen wählen. Sicherheitsprüfungen, Datenverwendung und unternehmensweite Kontrollmöglichkeiten klären und nachvollziehbar dokumentieren. Für geschäftssensible Arbeiten nur Lösungen und Zugänge einsetzen, die geeignete unternehmensweite Schutz- und Kontrollmechanismen bieten. Klar kommunizieren, dass nicht freigegebene Anwendungen („Schatten-KI“) und private Accounts für die geschäftliche Nutzung unzulässig sind und Mitarbeitenden geeignete Alternativen bereitstellen.

5. Gegen Manipulation schützen und Verlässlichkeit regelmäßig prüfen. Anweisungen in E-Mails, Webseiten und Dokumenten sowie direkt eingegebene manipulative Prompts dürfen einen Agenten nicht dazu bringen, seine Befugnisse zu überschreiten. Technische Beschränkungen, Freigaberegeln und Tests müssen zusammenwirken. Vor dem Einsatz, nach Änderungen und regelmäßig im Betrieb typische, schwierige und gezielt manipulierte Fälle in einer kontrollierten Umgebung testen. Feste Testanfragen und Anwendungsszenarien helfen, Abweichungen zu erkennen und nachzujustieren. Bei erhöhtem Risiko professionelle Sicherheitsprüfungen vorsehen.

6. Eingriff und Wiederherstellung vorbereiten. Aktionen nachvollziehbar protokollieren, Agenten kurzfristig stoppen und Zugriffe entziehen können. Festlegen, wer bei Zwischenfällen eingreift und wie fehlerhafte Buchungen, Nachrichten oder Datenänderungen erkannt und korrigiert werden. Notfallabläufe und die Wiederherstellung gesicherter Daten regelmäßig erproben. Aktionen, die sich nicht rückgängig machen lassen, erfordern besonders strenge Freigaben.

7. **Regeln und Kompetenzen laufend weiterentwickeln.** Verständliche Vorgaben, erreichbare Verantwortliche und geeignete Schulungen schaffen. Diese müssen verstehen, welche Daten sie freigeben dürfen, welche Befugnisse sie einem Agenten erteilen und wann menschliches Eingreifen erforderlich ist. Unsicherheiten und Auffälligkeiten sollten unkompliziert gemeldet werden können. Neue Fähigkeiten, Vorfälle und rechtliche Anforderungen regelmäßig zum Anlass nehmen, Regeln, Schutzmaßnahmen und Schulungsinhalte anzupassen.

Fazit

Die untersuchten Vorfälle stammen überwiegend aus besonderen Forschungsumgebungen. Sie zeigen, welche Fragen wir beim Einsatz handlungsfähiger KI stellen müssen. Sie sind kein Grund, Unternehmen vom Einsatz künstlicher Intelligenz abzuraten.

Im Gegenteil: KI kann touristische Unternehmen bei Beratung, Service und internen Abläufen erheblich unterstützen. Sie kann Mitarbeitende entlasten, Prozesse beschleunigen und neue Angebote ermöglichen. Diese Chancen zu nutzen, wird für die Wettbewerbsfähigkeit der Branche immer wichtiger. Gerade deshalb müssen die Verantwortlichen Sicherheit und Kontrolle von Anfang an mitdenken und ihre Regeln mit der Technologie weiterentwickeln.

Unser Ziel als Kompetenzkreis KI im Tourismus ist es, Unternehmen dabei zu unterstützen: **KI mutig und sinnvoll einsetzen, ihre Möglichkeiten laufend prüfen und Verantwortung für ihre konkreten Anwendungen übernehmen.** So entstehen aus technologischem Fortschritt verlässliche Angebote für Reisende und ein echter Vorteil für die Unternehmen.

Über den Kompetenzkreis KI im Tourismus

Künstliche Intelligenz verändert den Tourismus technologisch, organisatorisch und wirtschaftlich. Der Kompetenzkreis KI im Tourismus vereint unter fachlicher Leitung von Michael Buller und Alexander Mirschel derzeit 25 Expertinnen und Experten aus allen relevanten Bereichen der touristischen Wertschöpfungskette. Der Kompetenzkreis stärkt die Innovationsfähigkeit der Branche, macht Leuchtturmprojekte sichtbar und entwickelt tragfähige Lösungen für eine verantwortungsvolle und nachhaltige KI-Integration. Er bietet einen geschützten Raum für Austausch, Pilotprojekte und gemeinsame Positionierungen gegenüber Politik und Öffentlichkeit.

www.kompetenzkreis-ki.de

Kontaktdaten

Der Kompetenzkreis KI im Tourismus wurde initiiert und wird organisatorisch verantwortet von zwei zentralen Partnern: Verband Internet Reisevertrieb e. V. (VIR), Leonhardsweg 2, 82008 Unterhaching und Realizing Progress GmbH & Co. KG, Siedlerstraße 1c, 83607 Holzkirchen.

office@kompetenzkreis-ki.de, Internet www.kompetenzkreis-ki.de

Kompetenzkreis KI im Tourismus

Pressemeldungen im Internet: Die aktuellen Pressemeldungen stehen auch auf der Webseite des Kompetenzkreises KI im Tourismus unter „Publikationen” zum Download zur Verfügung.