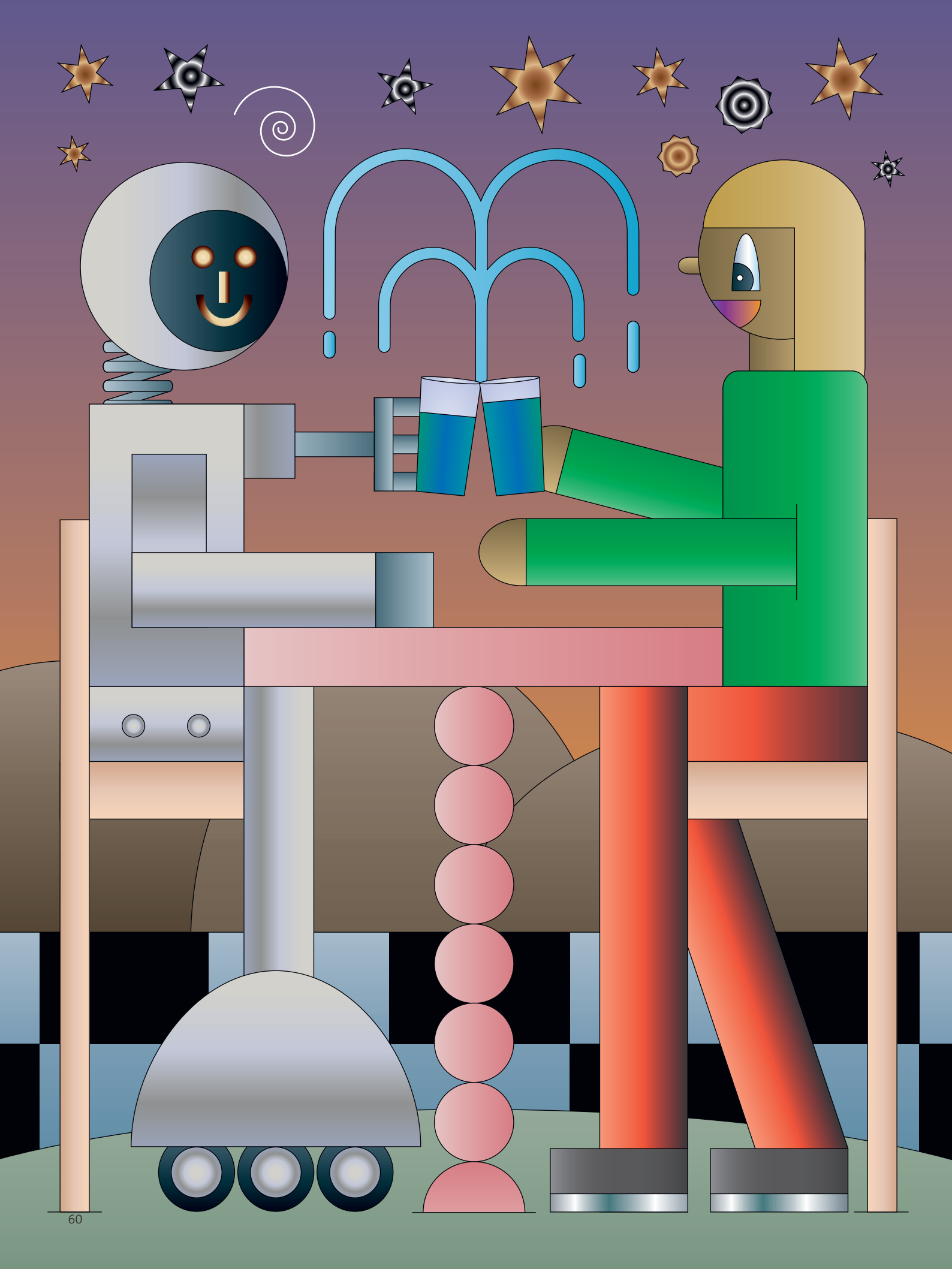


# Was bedeutet der Begriff VERTRAUEN in Bezug auf GENERATIVE KI?

Vertrauenswürdige KI gilt als zentrale Voraussetzung für ihren Einsatz. Doch was heißt es eigentlich, einer KI zu vertrauen? Der Beitrag zeigt, wie interdisziplinäre Forschung einen **gemeinsamen Vertrauensbegriff** entwickelt, der über einzelne Projekte hinaus nutzbar ist.

Von **Marvin Walczok, Celine Spannagl, Friederike Funk** und **Andreas Jungherr**

Generative KI-Systeme wie ChatGPT oder Copilot werden in immer mehr Arbeits- und Lebensbereichen genutzt. Sie erstellen in Sekundenschnelle Texte, Audios, Bilder und Videos und verändern so nachhaltig unterschiedlichste Anwendungsfelder: KI-generierte Inhalte finden sich in Nachrichtenformaten oder im politischen Wahlkampf, KI-Tutoren unterstützen beim Lernen, und zunehmend wird der Einsatz von KI auch in Hochrisikobereichen diskutiert, etwa bei der Urteilsfindung in Strafverfahren. Studien zeigen, dass die Nutzung von KI-Systemen weltweit kontinuierlich zunimmt, während gleichzeitig viele Menschen eine wachsende Skepsis gegenüber den Ergebnissen von KI, ihrer Zuverlässigkeit und ihren gesellschaftlichen Folgen äußern. In öffentlichen Debatten wird dieses Spannungsverhältnis häufig mit dem Begriff „Vertrauen" verbunden. Doch was ist damit genau gemeint? Können wir KI vertrauen? Und wann ist KI vertrauenswürdig?

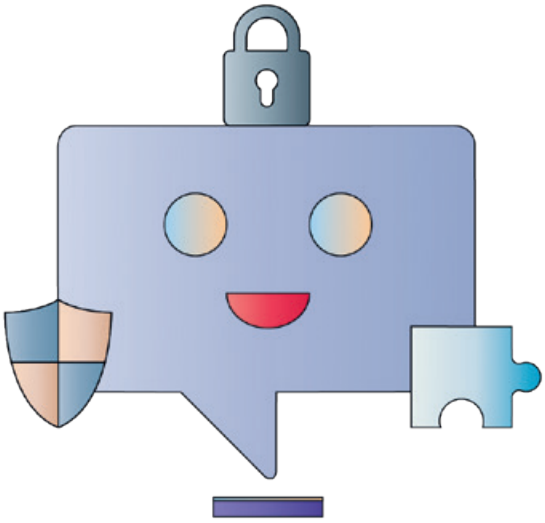
## Nutzung ist nicht gleich Vertrauen

Der Begriff „Vertrauen in KI" wird häufig verwendet, bleibt aber meist unscharf, gerade in der wissenschaftlichen Auseinandersetzung mit Mensch-KI-Interaktionen. Unterschiedliche Disziplinen verwenden ihn auf unterschiedliche Weise, häufig ohne ihn klar zu definieren. Ein einheitliches disziplinübergreifendes Verständnis fehlt, sodass interdisziplinäre Kooperationen vor der Herausforderung stehen, unterschiedliche Konzeptualisierungen und Methoden überbrücken zu müssen. Häufig wird Vertrauen im Kontext der KI mit der Nutzung eines KI-Systems gleichgesetzt, doch dies kann problematisch sein. Während die Nutzung ein beobachtbares Verhalten beschreibt, handelt es sich bei Vertrauen um eine innere Einstellung. Beide können stark voneinander abweichen: Die Nutzung von KI-Systemen muss keine Folge von hohem Vertrauen in KI sein.

## Vertrauen in Menschen – Vertrauen in KI

Ein sehr weit verbreitetes Vertrauensmodell stammt aus der Psychologie (Mayer et al., 1995). Es behandelt das zwischenmenschliche Vertrauen und unterscheidet dabei zwei Parteien: Vertrauensgeber (im Englischen: *trustor*) und Vertrauensempfänger (im Englischen: *trustee*). Wenn wir KI nutzen, nehmen wir die Rolle des Trustors ein, während die KI den Trustee repräsentiert. Das Modell definiert Vertrauen als die Bereitschaft des Trustors (Nutzende der KI), sich gegenüber den Handlungen eines Trustees (KI) vulnerabel zu zeigen, basierend auf der Erwartung, dass der Trustee wichtige Handlungen im Sinne des Nutzers ausführt – auch wenn dieser sie nicht kontrollieren kann. Besonders in risikoreichen Situationen spielt Vertrauen eine zentrale Rolle.

Illustration: Julian Litschko für Akademie Aktuell



KI-Nutzung erfordert kritisches Denken vor jeder Anwendung und ein Bewusstsein dafür, wo sie sinnvoll einsetzbar ist.

Entscheidend für die Vertrauenswürdigkeit des Trustees sind drei Faktoren: Kompetenz, Wohlwollen und Integrität. Die Kompetenz bezieht sich auf die wahrgenommenen Fähigkeiten des Trustees, eine spezifische Aufgabe zu erfüllen. Das Wohlwollen umfasst die wahrgenommenen positiven Absichten des Trustees gegenüber dem Trustor. Und Integrität beschreibt die wahrgenommene Einhaltung ethischer Grundsätze des Trustees. Generell gilt: Je höher diese Faktoren beim Trustee eingeschätzt werden, desto größer ist das Vertrauen des Trustors.

Eine Vielzahl von Studien hat nachgewiesen, dass dieses Modell des zwischenmenschlichen Vertrauens auch auf KI übertragbar ist. Nicht nur die wahrgenommene Kompetenz der KI, sondern auch das Wohlwollen und die Integrität bestimmen unser Vertrauen in KI und unser Vertrauensverhalten. Bei der Einschätzung der Vertrauenswürdigkeit ist die subjektive Wahrnehmung der KI entscheidend. Dies ist besonders wichtig hervorzuheben, denn KI hat kein Bewusstsein und kann aufgrund fehlender Intentionalität, Emotion und moralischer Verantwortlichkeit weder positive Absichten verfolgen noch mutwillig gegen ethische Grundsätze wie Fairness verstoßen. Trotzdem werden generativer KI häufig menschliche Eigenschaften zugesprochen, sodass manche Nutzende ihr Verhältnis zu Sprachmodellen wie ChatGPT als freundschaftlich oder therapeutisch betrachten.

Erweiterungen des Modells aus der Automatisierungsforschung zeigen zudem, dass für die erfolgreiche Nutzung von KI ein angemessenes Vertrauenslevel notwendig ist. Um die Chancen von KI zu nutzen und die Risiken zu minimieren, müssen Nutzende ihr Maß an Vertrauen proportional an die Fähigkeiten der KI anpassen. Ein unangemessenes Vertrauensniveau – sei es in Form von Über- oder Untervertrauen – kann negative Folgen haben, etwa durch unkritische Übernahme fehlerhafter Ergebnisse oder durch den Verzicht auf sinnvolle Unterstützung. Das Ziel sollte es daher nicht sein, Vertrauen pauschal zu stärken, sondern Nutzende darin zu unterstützen, KI differenziert und situationsabhängig einzusetzen. Dies erfordert kritisches Denken vor jeder Anwendung und ein Bewusstsein dafür, wo KI sinnvoll einsetzbar ist. Nur so können eine erfolgreiche Mensch-KI-Interaktion und eine Co-Kreation beider Parteien gelingen.

**Arbeitsmodell: Vertrauen in Mensch-KI-Interaktion**

**VERTRAUEN**

| **Funktion von Vertrauen** | **Individuelle Attributionen** | **Objektmerkmale der KI (Hinweisreize)** |
|---|---|---|
| Was soll Vertrauen leisten?<br>• Nutzung der KI<br>• Akzeptanz des Outputs<br>• Sicherheitsgefühl | Wie wird die KI subjektiv wahrgenommen?<br>• Kompetenz<br>• Wohlwollen<br>• Integrität<br><br>gemessen durch spezifische Skalen (kontext-/funktionsabhängig) | Welche Eigenschaften beeinflussen Wahrnehmung & Verhalten?<br>• Interfacegestaltung<br>• Erklärbarkeit<br>• Feedbackverhalten<br><br>typischerweise experimentell variiert |
| **bestimmt die Maße von** → | **Wahrnehmung & Einstellung der Nutzenden** | ← **beeinflussen je nach Variation** |

Illustrationen: Julian Litschko für Akademie Aktuell

## Vertrauen und Vertrauensverhalten

Kommen wir zurück zur Unterscheidung von Vertrauen und Vertrauensverhalten. In der Psychologie wird Vertrauen als eine Einstellung, beispielsweise gegenüber einem KI-System, verstanden, während Vertrauensverhalten eine resultierende Handlung beschreibt. Bei der Nutzung von Künstlicher Intelligenz kann sich das Vertrauensverhalten durch die Akzeptanz und Nutzung von KI-Systemen und KI-generierten Inhalten äußern. Diese Nutzung ist jedoch nicht zwangsläufig ein verlässlicher Indikator für vorhandenes Vertrauen. Unterschiedliche Faktoren können das beobachtbare Vertrauensverhalten beeinflussen, ohne dass sie sich gleichzeitig auch auf das Vertrauen an sich auswirken.

Ein Beispiel: Nehmen wir an, Sie erhalten von Ihrer Führungskraft eine wichtige, für Sie neue und zugleich komplexe Aufgabe. Dies könnte die statistische Analyse eines Datensatzes, das Drehen eines Werbevideos oder das Erstellen eines Quartalsberichts sein. Die Aufgabe müssen Sie innerhalb weniger Stunden erledigen, Sie haben also nicht genügend Zeit, sich vollständig einzuarbeiten. Um den Termin einzuhalten, nutzen Sie ChatGPT und geben die vorgeschlagene Lösung mit einigen Änderungen ab, sind sich aber unsicher, ob die KI Ihre Aufgabe gut bearbeitet hat und Ihre Lösung korrekt ist. In diesem Beispiel zeigen Sie Vertrauensverhalten trotz geringen Vertrauens in ChatGPT, bedingt durch Zeitdruck und fehlende Expertise.

## Impulse für die Forschung

Eine einheitliche Definition von Vertrauen ist für die interdisziplinäre Forschung zu Künstlicher Intelligenz von zentraler Bedeutung. Jedoch kann Vertrauen abhängig vom konkreten Forschungsprojekt unterschiedliche Rollen einnehmen. Um die Vergleichbarkeit zwischen einzelnen Projekten zu erhöhen, empfehlen wir Forschenden daher, genau zu spezifizieren, ob es jeweils um die wahrgenommene Vertrauenswürdigkeit der KI und/oder ein resultierendes Vertrauensverhalten geht. Damit Sie herausfinden, welche Konzepte relevant sind, beantworten Sie die folgenden Fragen:

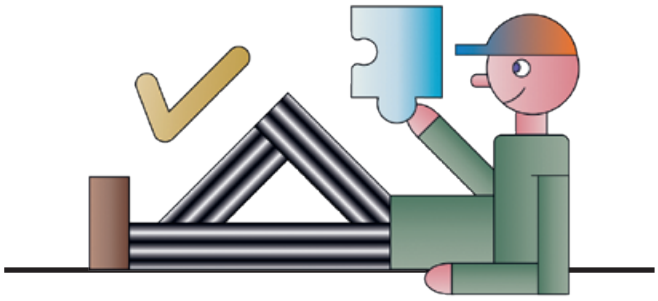
**1. Funktion:**
Welche Funktion soll das Vertrauen in KI im untersuchten Kontext erfüllen? Soll es die Nutzung der KI oder die Akzeptanz ihrer Ergebnisse fördern?

**2. Individuelle Einschätzung:**
Wie schätzen Nutzende die Facetten der Vertrauenswürdigkeit (Kompetenz, Wohlwollen, Integrität) der KI ein? Welche Messinstrumente werden verwendet? Bezieht sich das Vertrauen auf das System, den Prozess, die Interaktion oder das Endprodukt?

**3. Objektebene:**
Welche Eigenschaften der KI beeinflussen die wahrgenommene Vertrauenswürdigkeit? Wie wirken sich Veränderungen dieser Eigenschaften auf das Vertrauen in KI und das Vertrauensverhalten aus?

## Fazit

Interdisziplinäre Forschung profitiert von einem gemeinsamen Verständnis von Vertrauen, vertrauenswürdiger KI und daraus resultierendem Verhalten, um die Bedingungen erfolgreicher Mensch-KI-Interaktion besser zu verstehen. Entscheidend ist nicht blindes Vertrauensverhalten, sondern ein situationsangemessenes, reflektiertes Vertrauen. Letztendlich ist eine KI vertrauenswürdig, wenn sie sich in der Praxis des ihr entgegengebrachten Vertrauens würdig erweist.



**Dr. Marvin Walczok**
ist wissenschaftlicher Mitarbeiter am Lehrstuhl für Sozialpsychologie mit Schwerpunkt Rechtspsychologie der LMU München. Im Schwerpunkt „Mensch und generative KI: Trust in Co-Creation" des Bayerischen Forschungsinstituts für Digitale Transformation (bidt) der BAdW erforscht er Vertrauen in generative KI im Kontext des Rechtssystems.

**Celine Spannagl**
ist wissenschaftliche Referentin in der Abteilung Forschung am bidt. Sie forscht im bidt-Forschungsschwerpunkt und am Projekt „Ethische Implikationen hybrider Teams aus Mensch und KI-System" zum Thema Erklärungen und Vertrauen in der Mensch-KI-Interaktion.

**Prof. Dr. Friederike Funk**
ist Professorin für Sozialpsychologie mit Schwerpunkt Rechtspsychologie an der LMU München. Sie forscht über subjektives Gerechtigkeitserleben und Personenwahrnehmung im rechtlichen Kontext. Im bidt-Schwerpunkt „Mensch und generative KI: Trust in Co-Creation" untersucht sie psychologische Determinanten von Vertrauen in die Co-Kreation mit generativer KI im Rechtssystem.

**Prof. Dr. Andreas Jungherr**
ist Inhaber des Lehrstuhls für Politikwissenschaft, insbesondere Digitale Transformation, an der Universität Bamberg und Mitglied des bidt-Direktoriums. Er erforscht, wie Digitalisierung und KI politische Prozesse, öffentliche Kommunikation und gesellschaftliche Dynamiken verändern. Im Rahmen des bidt verantwortet er interdisziplinäre Projekte zu generativer KI, politischer Meinungsbildung und Desinformation.